MEMAHAMI KERANGKA *DATA ANALYTICS* PADA SITUS COVID19.GAMABOX.ID DAN MELAKUKAN *EXPLORATORY DATA ANALYSIS* PADA DATASET COVID19.GAMABOX.ID

Hendri Kurniawan P.
19/448706/PPA/05789

A. Landasan Teori

Data analytics adalah proses ekstraksi pengetahuan yang mendalam dari *data collection* yang besar untuk membantu seseorang atau organisasi dalam membuat keputusan yang lebih baik. *Data analytics* menggunakan algoritma *machine learning*, *artificial intelegence*, statistik dan *natural language processing* untuk menemukan pola dalam data dengan memanfaatkan *data visualization tools* untuk membuat pola tersebut dapat dipahami oleh users.

Berdasarkan tujuannya, *data analytics* dibedakan menjadi 3 jenis yaitu: *descriptive statistics, predictive analytics, prescriptive analytics*. *Descriptive statistics* digunakan untuk memahami apa yang telah terjadi ***(what happened)*** dan kenapa ***(why)*** hal itu terjadi dalam bentuk *summary* dari *data collections* tersebut. *Predictive analytics* digunakan untuk melakukan analisis kondisi yang akan terjadi pada masa yang akan datang ***(what will happen next)***. Sedangkan *prescriptive analytics* digunakan untuk menganalisis apa yang harus dilakukan pada kondisi yang akan terjadi tersebut ***(what to do)***[1].

Gambar 1 Level *data analytics*

Sebagai contoh dalam kajian riwayat belanja pelanggan, permasalahan yang dapat diselesaikan pada level *analytics descriptive* adalah[2]:

- siapa 10 pelanggan dengan belanjaan terbanyak?
- di hari apa saja pelanggan ramai membeli?
- berapa rata-rata nilai belanjaan tiap pelanggan?
- apakah ada perbedaan jumlah belanjaan di hari libur dan hari kerja?
- adakah hubungan antara hari libur dan jumlah belanjaan?

Kemudian beberapa pertanyaan yang dapat dijawab pada level *predictive analytics* adalah:

- berapa penjualan harian untuk 1 bulan ke depan?
- pelanggan mana saja yang akan membeli lebih banyak jika diberikan diskon?
- pelanggan mana saja yang akan berhenti membeli?

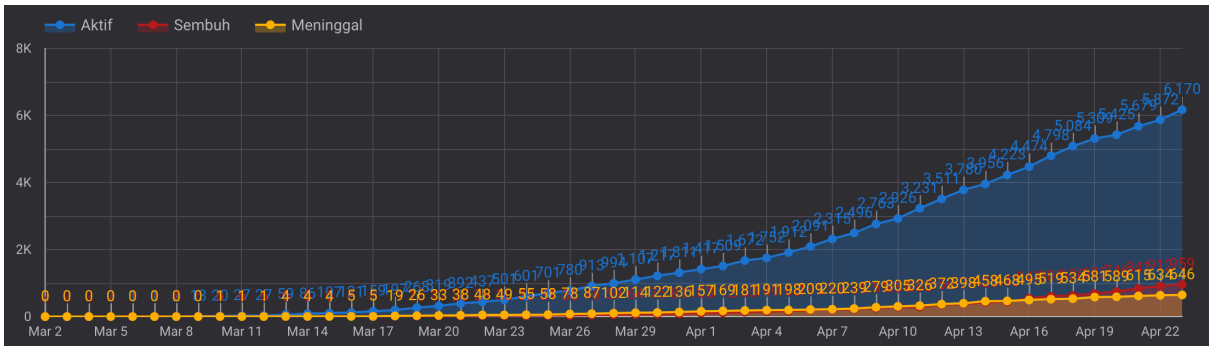
Sedangkan untuk prescriptive analytics yaitu:

- berapa % diskon untuk pelanggan yang rutin belanja tiap minggu supaya nilai belanjaannya meningkat?
- produk apa yang sebaiknya ditawarkan kepada pelanggan yang merupakan pasangan baru menikah?
- dimana sebaiknya meletakkan suatu produk supaya dapat terjual lebih banyak?

B. Pembahasan

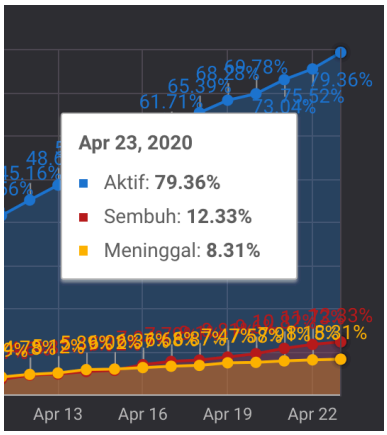
Pendahuluan

Situs covid19.gamabox.id adalah salah satu laman yang menyajikan data tentang perkembangan terkini kondisi pandemi virus covid-19. Dalam penyajian datanya situs tersebut menggunakan kerangka *data presentation*. *Data presentation* digunakan untuk memahami *landscape* data dengan pendekatan visualisasi yang menarik dalam tingkatan analisis deskriptif, yaitu menjelaskan apa yang telah atau sedang terjadi. Terlihat pada gambar 2.



Gambar 2 Grafik pertumbuhan kasus Covid-19 perhari

Dari gambar tersebut terlihat bahwa kecenderungan kasus tiap harinya meningkat. Peningkatan signifikan terjadi setelah tanggal 4 April. Dalam kurun waktu 9 hari kasusnnya meningkat sampai 2000 kasus. Hingga tanggal 23 April 2020 persentase kasus sembuh sebanyak 12.33% dan meninggal mencapai 8.31%. terlihat pada gambar 3.



Gambar 3 Persentase jumlah kasus

Sajian tersebut akan menjadi lebih menarik jika masuk ke wilayah *predictive analytics.* Sederhananya saja dengan mengetahui rata-rata laju pertambahan kasus perharinya maka dapat diestimasi lonjakan kasus tertinggi akan terjadi kapan. Apabila ingin lebih *advance* lagi dengan menggunakan algoritma *machine learning* dapat diprediksi kapan pandemi ini akan berakhir.

Selain itu jika data tersebut dikombinasikan dengan data ketersediaan tenaga atau fasilitas kesehatan dapat dihitung kemampuan tiap daerah dalam menghadapi *pandemic* ini. Sehingga dapat dijadikan pertimbangan dalam penyusunan regulasi. Hal tersebut akan menjadi lebih bermanfaat daripada sajian yang sifatnya hanya informatif saja

Daftar Faskes

	Nama Faskes	Jumlah Nakes	Alamat
1.	Rs dr. Sadikin kota pariaman	>150	Rs dr. Sadikin kota pariaman Jl.nostalgia kp. Bar...
2.	BLUD RS KONAWE	>150	BLUD RS KONAWE Jl. P Diponegoro No. 301 Un...
3.	Puskesmas samata kab. Gowa.sul sel	51-100	Puskesmas samata kab. Gowa.sul sel Jl Mustaf...
4.	Rumah Sakit Kristen Lindimara waingap...	>150	Rumah Sakit Kristen Lindimara waingapu sumb...
5.	UPT BLUD PUSKESMAS GANGGA	101-150	UPT BLUD PUSKESMAS GANGGA Jalan Raya Ju...
6.	Puskesmas Maratua	20-50	Puskesmas Maratua Jl. Tan Ten Siang, Teluk Ha...
7.	Puskesmas Paloh	51-100	Puskesmas Paloh Desa nibung Sambas Kalima...
8.	Puskesmas Pandean	20-50	Puskesmas Pandean Jl. Raya Trenggalek-Pangg...
9.	Puskesmas duduksampeyan	51-100	Puskesmas duduksampeyan Jalan raya duduks...
10.	rsu budirahayu pekalongan	>150	rsu budirahayu pekalongan Jl. Barito No.5, Duku...

Gambar 4 Tabel ketersedian tenaga kesehatan

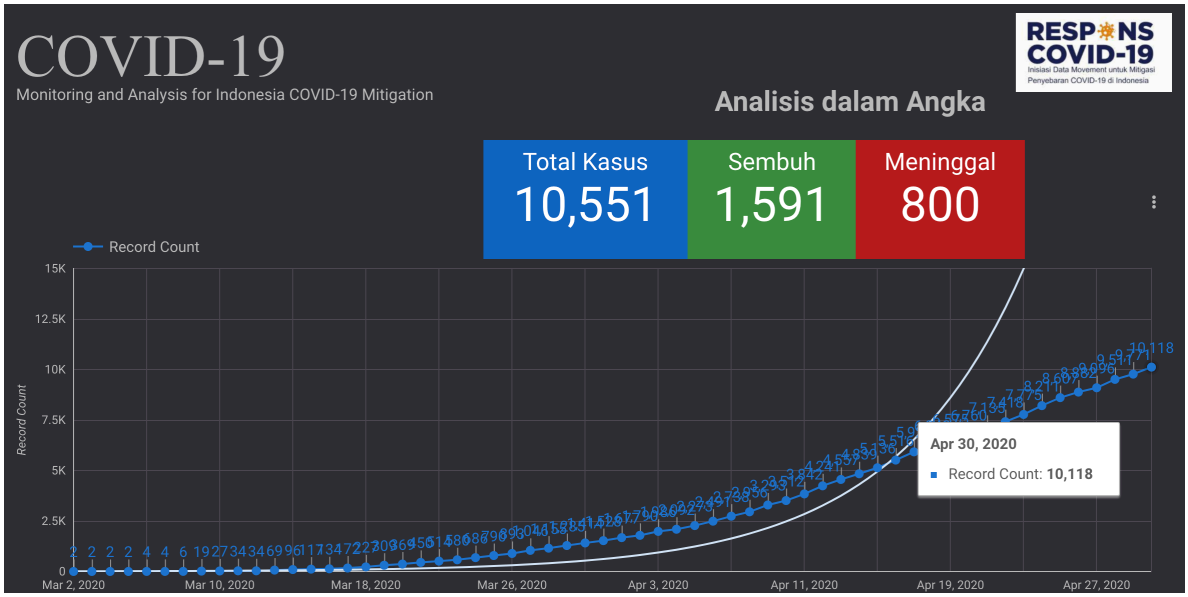
Hasil Analisis Deskriptif

Dengan menggunakan dataset yang tersedia dapat dilakukan analisis sederhana menggunakan metode statistika deskriptif.

	count	mean	std	min	25%	50%	75%	max
Aktif	54.0	144.555556	102.610578	1.0	65.75	114.5	224.75	352.0
Meninggal	47.0	16.829787	13.133626	1.0	8.00	12.0	22.50	60.0
Sembuh	48.0	31.687500	35.420367	1.0	4.00	19.0	46.25	144.0

Gambar 5 Statistika deskriptif pada data covid19.gamabox.id

Salah satu hipotesa sederhana yang dapat diambil dari data tersebut adalah dengan menghitung rata-rata kasus meninggal ditambah kasus sembuh dan dengan menghentikan laju pertambahan kasus positif hingga 0, yang dapat diasumsikan dari pemberlakuan *lockdown* secara total pada setiap daerah maka dengan menyisakan kasus positif yang ada dapat dihitung lama (dalam hari) kapan pandemi akan berakhir.



Gambar 6 Analisis dalam angka

Source: *http://covid19.gamabox.id/analysis diakses tanggal 2 Mei 2020*

Anggap kasus terjadi mulai tanggal 1 Maret 2020 hingga 30 April 2020 yaitu berlangsung selama 61 hari. Jika total kasus positif mencapai 10.551, sembuh 1.591 dan meninggal 800 sehingga kasus aktif masih 8200 kasus.

Dengan menghitung rata-rata kasus sembuh (dengan pembulatan) mencapai 26 orang sembuh dan 13 orang meninggal perhari maka **8200 kasus akan selesai pada 8200/(26+13) = 210 hari (+/- 7 bulan)**. Tentunya bagi kita rata-rata kematian yang mencapai 13 perhari adalah angka yang buruk. **Sehingga perlu ada upaya untuk menekan angka kematian dan menambah angka sembuh.**

Bagi pemerintah ataupun stakeholder yang terkait, waktu *lockdown* 7 bulan adalah hal yang sulit. **Sehingga perlu menyusun strategi atau kebijakan yang tepat, misalnya saja dengan meningkatkan kualitas penanganan kasus positif untuk menambah angka kesembuhan maka perlu adanya tambahan dokter dan alat kesehatan yang memadai**. Tambahan dokter bisa didatangkan dari negeri seberang. Pengadaan alat bisa diwujudkan dengan penambahan anggaran. Hal tersebut dapat dihitung jika kita dapat mengestimasi kapan pandemi akan berakhir.

Selain itu dengan mengetahui kapan pandemi akan berakhir pemerintah atau lembaga kemanusiaan **dapat mengestimasi kebutuhan masyarakat selama *lockdown*** berlangsung dan institusi lain seperti perkantoran, universitas, sekolah dan yang lainnya juga dapat menyesuaikan kegiatannya.

Hasil Analisa Berbasis *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) dilakukan dengan menggunakan *tools programming* Python Jupyter Notebook. EDA adalah proses untuk memahami data sebelum menentukan model *machine learning* yang akan digunakan. Beberapa proses yang telah dilakukan yaitu:

1. Meneliti kelengkapan data

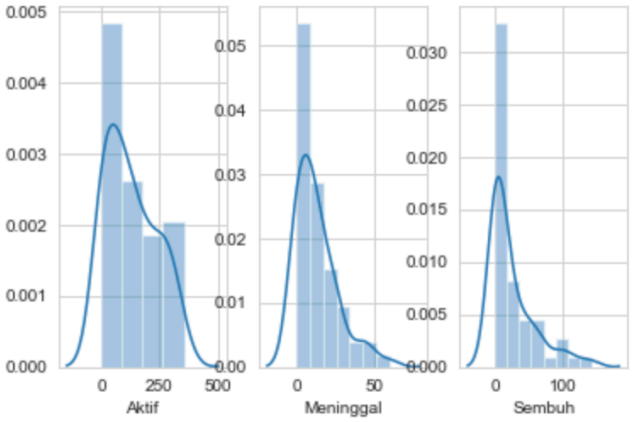
```
In [12]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61 entries, 0 to 60
Data columns (total 3 columns):
Aktif        54 non-null float64
Meninggal    47 non-null float64
Sembuh       48 non-null float64
dtypes: float64(3)
memory usage: 1.5 KB
```

Gambar 7 Proses validasi kelengkapan data

Pada gambar 7 terlihat tipe data, jumlah baris dan keterisian tiap kolomnnya. Terlihat bahwa terdapat variabel yang belum lengkap. Kolom "Aktif" dari total 61 baris data terdapat 7 baris yang kosong dan semua variabel bertipe float. Baris yang kosong (tidak ada nilainya/NaN) akan di isi dengan nilai 0.

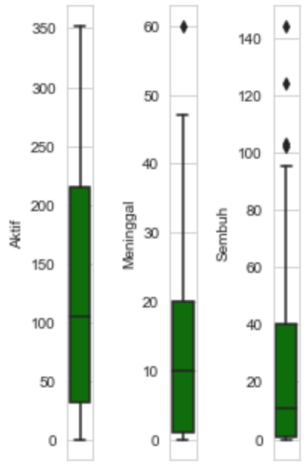
2. Melakukan perhitungan statistik deskriptif untuk menghitung nilai rata-rata, quartil 1, median dan quartil 3, standar deviasi serta nilai minimal-maksimal. Hasilnya dapat diamati pada gambar 5 menunjukkan bahwa nilai rata-rata lebih besar dari pada nilai mediannya. sehingga distribusi data menjadi tidak normal. kecenderungan kurvanya akan melenceng ke kanan (skewed negatif).
3. Visualisasi data menggunakan *plot distribution*



Gambar 8 Diagram plot distribusi

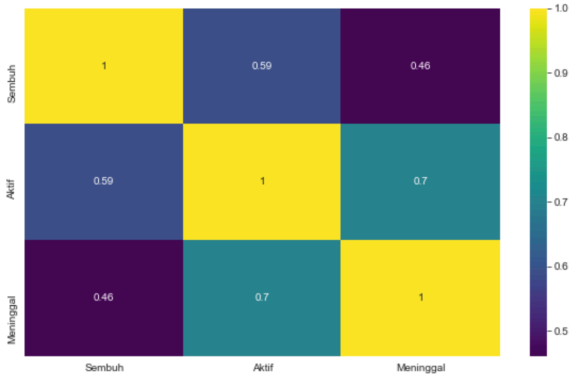
Gambar 8 menunjukkan bahwa data yang digunakan bentuknnya tidak berdistribusi normal. Sesuai dengan deskripsi pada nomor 2.

4. Meneliti adanya anomali menggunakan diagram *blox plot.* Terlihat pada gambar 9 adanya anomali data pada variabel meninggal dan sembuh.



Gambar 9 Diagram box plot untuk menunjukkan anomali data pada variabel meninggal dan sembuh

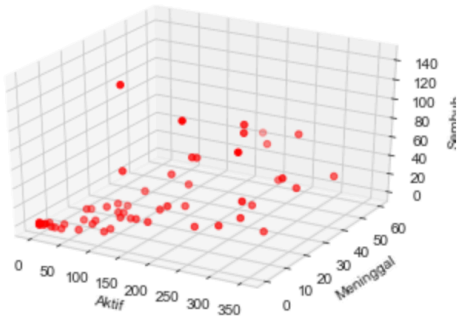
5. Melakukan analisis korelasi variabel. Analisis korelasi menjelaskan ada atau tidaknya hubungan antar dua variabel. Nilai Korelasi bisa positif atau negatif atau lemah. Korelasi positif yang artinya jika penambahan pada nilai X maka bertambah juga nilai Y. Korelasi negatif menjelaskan hubungan setiap kenaikan nilai X maka ada penurunan pada nilai Y. Korelasi yang lemah menjelaskan dua variabel ini tidak ada hubungannya sama sekali. Biasanya korelasi dikatakan sangat kuat jika nilainya melebihi 0.7 jika kurang dari tersebut korelasi antar dua variabel tersebut lemah



Gambar 10 Visualisasi analisis korelasi

Terlihat dari diagram 10, tidak semua variabel berkorelasi kuat. Namun ada 2 variabel yang menunjukkan korelasi cukup kuat yaitu variabel "aktif" dan "meninggal". Jika akan dilakukan analisis regresi dapat menggunakan variabel tersebut.

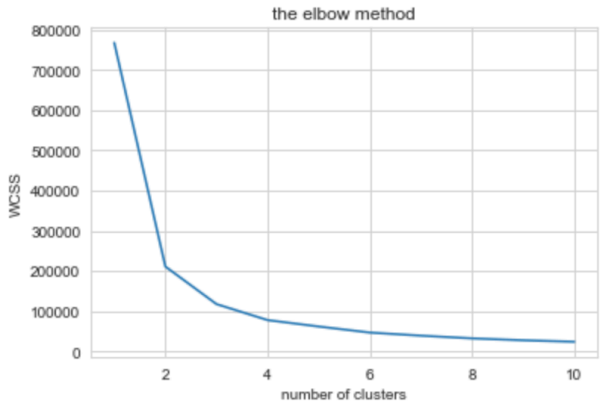
6. Melakukan analisis klaster. Karena data yang diamati tidak memiliki label maka dilakukan analisis klaster untuk mengelompokkan data berdasarkan tingkat similaritasnya. Persebaran data dapat dilihat pada gambar 11.



Gambar 11 Persebaran data kasus positif covid-19

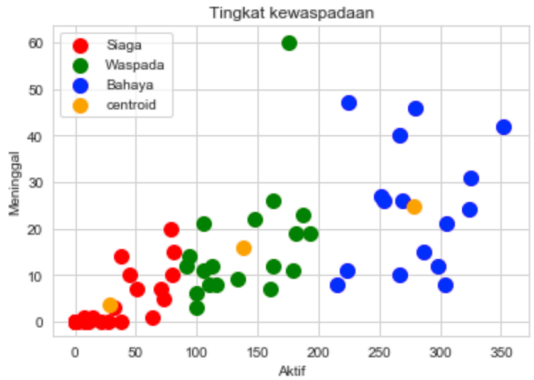
Berdasarkan proses analisis diatas maka untuk menarik informasi yang mendalam dari data yang tersedia dapat digunakan algoritma K-Means untuk melakukan proses klastersasi pada data. **Tujuannya adalah untuk mengklaster status kegawat daruratan covid-19 berdasarkan angka kasus positif (aktif, meninggal, sembuh)**.

Proses klasterisasi menggunakan *elbow method* untuk menentukan jumlah klaster yang terbaik. Hasilnya adalah terbentuk klaster optimal sebanyak 3 klaster ditunjukkan pada gambar 12.



Gambar 12 Perhitungan nilai jarak minimal per klaster

Hasil klaster ditunjukkan pada gambar 13. Yaitu klaster yang dilabeli dengan status “Siaga”, “Waspada” dan “Bahaya”.



Gambar 13 Hasil klaster yang ditunjukkan dalam 2 dimensi

Berikut nilai tiap klasternya disajikan pada tabel 1.

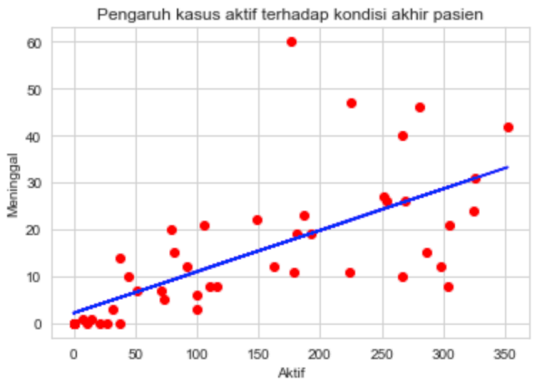
Tabel 1 Rentang nilai status kewaspadaan

Status	Nilai Awal	Nilai Akhir
Siaga	Aktif:1, Sembuh: 1	Aktif: 81, Meninggal: 15, Sembuh: 13
Waspada	Aktif: 92, Meninggal: 12, Sembuh: 9	Aktif: 193, Meninggal: 19, Sembuh: 71
Bahaya	Aktif: 215, Meninggal: 8, Sembuh: 124	Aktif: 352, Meninggal: 42, Sembuh: 42

Diharapkan dengan adanya status kewaspadaan dapat menjadi pertimbangan dalam menyusun regulasi oleh pemerintah atau institusi lainnya dan memudahkan masyarakat dalam menyimpulkan kondisi.

Analisis klastering juga bisa digunakan untuk mendeteksi anomali data, karena jika kita lihat pada gambar 13 terdapat titik point yang berada jauh dari klasternya. Hal ini bisa dijadikan modal untuk investigasi lebih dalam terhadap kasus tersebut. Apa yang menyebabkan anomali dan kenapa bisa terjadi.

Selain analisis klastering, dapat dilakukan juga analisis regresi untuk mengetahui keterkaitan antar 2 variabel. Berdasarkan hasil pengamatan korelasi variabel terdapat 2 variabel yang memiliki korelasi positif yaitu variabel “aktif” dan “meninggal”. Hasil analisis regresi dapat dilihat pada gambar 14.



Gambar 14 Analisis keterkaitan kasus covid-19 “aktif” dan “meninggal”

Dari analisis tersebut kita dapat memprediksi angka kematian berdasarkan kasus yang masih aktif dengan persamaan linear dari garis biru tersebut.

PUSTAKA

[1] Darono Agung, Pembelajaran Business Analytics dan Big Data Dalam Pendidikan Ekonomi dan Bisnis, National Seminar on Accounting and Finance, Oktober 2016.

[2] http://www.datapublik.com/2016/02/3-kategori-analisis-data.html

[3] http://skj.mipa.ugm.ac.id/2020/04/20/4-big-data-analytic-descriptive-and-perscriptive/