Nama : Randy Leonard
NIM : 2201753826
Kelas : LD-01

1. Patient-meta data

Analisis cara 1

Covid 19 Suspected Movement Case

This dataset describes about Covid19 Case. This dataset is from surveyor that gather data from covid19 suspected person in movement/mobility before they are quarantined.

Data Source : Apache Hive

Name	Descriptions	Data Type	More Information
Name	Name of suspected person	String	
Gender	Gender	String (Enum)	Male, Female
Age	Age of patient	Int	
PersonStatus	Covid19 suspected category	String(Enum)	*PDP, ODP*
PlaceName	Name of place	String	
PlaceType	Place descriptions	String (Enum)	Restaurant, Office, Public, Market, Private
Duration	Duration suspected person take a break in this place (minutes)	Int	
Latitude	Coordinate of place	Double	
Longitude	Coordinate of place	Double	
VisitedOn	Suspected person make visited in date	DateTime	

Micro array-data

Hospital Covid19 Case

This dataset describes about the status of covid19 patient in care at hospital. This data is updated and inserted daily.

Data Source : csv (HDFS)

Name	Descriptions	Data Type	More Information
Name	Name of patient	String	
Gender	Gender	String(Enum)	Male, Female
Age	Age of patient	Int	
HospitalID	Hospital ID based on master data	Int	
HospitalName	Hospital Name	String	
IsPositive	Covid19 positive status	Boolean	
PatientCovidStatus	Covid19 suspected category	Enum (string)	*ODP, PDP*
IsTested	Patient is tested for swab test	Boolean	
TestedDate	Patient swab test date	DateTime	
PatientStatus	Patient status in hospital	String(Enum)	PassedAway, InCare, Healed
CreatedOn	Created record	DateTime	

Hospital Medical Item Needs

The dataset describes about medical item that hospital needs.
Data Source : csv (HDFS)

Name	Descriptions	Data Type	More Information
HospitalID	Hospital ID	Int	
HospitalName	Hospital Name	String	
MedicalItemName	Medical item that hospital need	String	Surgical Mask, Hazmat Suite, Disinfectant
Amount	Amount item that hospital needs	Double	
CreatedOn	Covid19 positive status	Boolean	

Hospital meta-data

Select hospital with a particular set of functions and join hospital meta-data with patient meta-data and microaway data

PatientID	PersonStatus	HospitalID	PatientStatus	PlaceType
0	ODP	1	InCare	Public
1	PDP	1	PassedAway	Market
2	PDP	1	Healed	Restaurant
3	ODP	1	Healed	Market
4	PDP	1	PassedAway	Public
5	PDP	1	Healed	Market
6	ODP	1	InCare	Office
0	ODP	2	InCare	Market
1	PDP	2	PassedAway	Public
2	ODP	2	Healed	Market
3	PDP	2	PassedAway	Market
4	ODP	2	InCare	Offiec
5	PDP	2	Healed	Public
6	PDP	2	Healed	Market

PIVOT

PIVOT

Pivot the table in previous step to get expression values for each types of hospit for each patient

PatientID	1	2
0	InCare	InCare
1	PassedAway	PassedAway
2	Healed	Healed
3	Healed	PassedAway
4	PassedAway	Healed
5	Healed	PassAway
6	InCare	InCare

Select patientID, person status,place name,duration from patient meta-data

PatientID	PersonStatus	PlaceType
0	ODP	Public
1	PDP	Market
2	PDP	Restaurant
3	ODP	Market
4	PDP	Public
5	PDP	Market
6	ODP	Office

JOIN

JOIN

PatientID	PersonStatus	PlaceName	PatientID	1	2
0	ODP	Public	0	InCare	InCare
1	PDP	Market	1	PassedAway	PassedAway
2	PDP	Restaurant	2	Healed	Healed
3	ODP	Market	3	Healed	PassedAway
4	PDP	Public	4	PassedAway	Healed
5	PDP	Market	5	Healed	PassAway
6	ODP	Office	6	InCare	InCare

Select patients with some person status and join results with the micro-array table

PatientID	PersonStatus	HospitalID	PatientStatus
3	PDP	0	PassedAway
3	PDP	1	InCare
3	PDP	2	PassedAway
3	PDP	3	Healed
3	PDP	4	InCare
3	PDP	5	PassedAway
3	PDP	6	InCare
6	PDP	0	Healed
6	PDP	1	PassedAway
6	PDP	2	InCare
6	PDP	3	PassedAway
6	PDP	4	Healed
6	PDP	5	PassedAway
6	PDP	6	InCare

Pivot the table to get the expression values for all hospital for each patient

Patientid	1	2	3	4	5	6
1	PassedAway	InCare	PassedAway	Healed	InCare	Healed
2	InCare	PassedAway	InCare	Healed	PassedAway	InCare
6	PassedAway	Healed	PassedAway	InCare	Healed	InCare

Compute the correlation between the exoression levels of all paires of hospital.
Corelation Matrix

1	Hospital-1	Hospital-2	Hospital-3	Hospital-4	Hospital-5	Hospital-6
Hospital-0	1	-0,47259	-0,8211	-0,0031	0,62313	0,.99999
Hospital-1	-0,47252	1	-0,8232	0,27362	-0,6213	0,82721
Hospital-2	-0,738213	0,313121	1	-0,5432	-0,9971	0,99991
Hospital-3	-0,4561	-0,84576	-0,37361	1	0,62231	-0,81313
Hospital-4	0,87622	-0,82122	0,462357	0,65510	1	-0,87555
Hospital-5	-0,9999	0,981200	-0,98888`	0,76212	0,898711	1
Hospital-6	-0,75212	0,520000	0,97000	0,99997	0,994608	1

Use this table to Train Linear Regression models to predict

2.Analisis Cara 2
First Analytics
use data sets of Covid 19 Suspected Movement Case (Name, Gender, Age, PersonStatus, PlaceName, PlaceType, Duration, Latitude, Longitude, VisistedOn), data sets of Hospital Covid19 Case (Name, Gender, Age, HospitalID, HospitalName, IsPositive, PatientCovidStatus, IsTested, TestedDate, PatientStatus, CreatedOn) and data sets of CCTV Data (ID, Latitude, Longitude, VideoData, CreatedOn). Based on the data sets that I chose, we can calculate the probability of the place that has the most likely area Covid 19 will appear. By using these three data sets, we can find each suspected case and hospital cases by using data of longitude and latitude, and also we can view the video through CCTV to monitor all area. With longitude and latitude, we can predict areas that may be infected by covid 19 and can mark the place as a dangerous place. Even though it is only a prediction, but first we will test to measure the probability of the prediction areas
Example of covid19 suspected case:
- Fepri Putra, Male, 30, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10
- Ani, Female, 52, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10

Example of Hospital covid19 case:
- Fepri Putra, Male, 30, 1, RS.Sulianti Saroso, True, PDP, True, 10-05-2020 10:10:10, InCare, 10-05- 2020 10:10:10

Second Analytics
use data sets of Hospital Covid19 Case (Name, Gender, Age, HospitalID, HospitalName, IsPositive, PatientCovidStatus, IsTested, TestedDate, PatientStatus, CreatedOn) and data sets of Hospital Medical Item Needs (HospitalID, HospitalName, MedicalItemName, Amount, CreatedOn) Based on the data sets of Hospital Covid19 Case and data sets of Hospital Medical Item Needs, we can calculate and count the medical items that needed to use for patients and doctors so for the entire treatment process can be carried out smoothly when the cases is increasing suddenly.
Example of covid19 suspected case:

- Fepri Putra, Male, 30, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10
- Ani, Female, 52, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10

Example of Hospital Medical Item Needs:
- 1, RS.Sulianti Saroso, Surgical Mask, 1000, 10-05-2020 10:10:10
- 1, RS.Sulianti Saroso, Hazmat Suite, 1000, 10-05-2020 10:10:10

Third Analytics
I will use data sets of Covid 19 Suspected Movement Case (Name, Gender, Age, PersonStatus, PlaceName, PlaceType, Duration, Latitude, Longitude, VisistedOn) and data sets of Social Assitance Distributions (SubDisctrictID, SubDistrictName, DistrictID, DistrictName, ProvinceID, ProvinceName, TotalAllocation, DateAllocation, Institution). Based on data sets from Covid 19 Suspected Movement Case and data sets from Social Assistance Distributions, we can calculate the allocation costs needed for each areas based on data set of Social Assitance Distributions which is subdistrict, district, and province and also data set of Covid 19 Suspected Movement Case, which is show the covid 19 cases based on latitude and longitude. With these both of data sets, all cost can be estimated clearly so that all allocated funds can be given to the affected communities.
Example of covid19 suspected case:
- Fepri Putra, Male, 30, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10
- Ani, Female, 52, ODP, Ps.TanahAbang, Market, 120, 106.8097425, -6.1890043, 10-05-2020 10:10:10

Example Social Assistance Distribution
- DKI1, Keb.Baru, DKI1, JakartaSelatan, 1, DKI Jakarta, 10000000, 10-05-2020 10:10:10, ABC.PT
- DKI1, Keb.Baru, DKI1, JakartaSelatan, 1, DKI Jakarta, 10000000, 10-05-2020 10:10:10, HambaTuhan

Spark implementation for predicting status of covid19 patient in care at hospital using regression model

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.regression import LinearRegressionWithSGD
from pyspark.mllib.regression import LinearRegressionModel
sc = SparkContext(appName="App")
sqlContext = SQLContext(sc)
hospital = sc.textFile('/home/ubuntu/HospitalMetaData-10-10.txt')
header = sp.first() #extract header
hospital = .filter(lambda x:x !=header)
hparts = hospit.map(lambda l: l.split(", "))
```

```
hospitalframe = hparts.map(lambda p: Row(hospitalid=int(p[0]),
target=int(p[1]), position = long(p[2]),
length=int(p[3]), function=int(p[4])))
schemaHospit = sqlContext.createDataFrame(hospitalframe)
schemaHospit.registerTempTable("hospitals")
patients = sc.textFile('/home/ubuntu/PatientMetaData-10-10.txt')
header = patients.first() #extract header
patients = patients.filter(lambda x:x !=header)
pparts = patients.map(lambda l: l.split(", "))
patientsframe = pparts.map(lambda p: Row(patientid=int(p[0]),
age=int(p[1]), gender=int(p[2]),
personstatus=int(p[3]),placetype =int(p[4]),
placeType = int(p[5])))
schemaPatients = sqlContext.createDataFrame(patientsframe)
schemaPatients.registerTempTable("patients")
geo = sc.textFile('/home/ubuntu/GEO-10-10.txt')
header = geo.first() #extract header
geo = geo.filter(lambda x:x !=header)
geoparts = geo.map(lambda l: l.split(", "))
geoframe = geoparts.map(lambda p: Row(hospitalid=int(p[0]),
patientid=int(p[1]), patientstatus = float(p[2])))
schemaGEO = sqlContext.createDataFrame(geoframe)
schemaGEO.registerTempTable("geo")
g = sqlContext.sql("SELECT p.patientid, p.personstatus,
e.hospitalid, e.hospitalname, p.patientstatus FROM
hos AS g, patients AS p, geo AS e
WHERE g.function < 300 AND
g.patientid = e.patientid
AND p.patientid = e.patientid")
h.registerTempTable("responses")
h2=g.groupBy('patientid').pivot('hospitalid').sum('PatientStatus')
h2.registerTempTable("gen")
h3 = sqlContext.sql("SELECT patientid,age,name,
personStatus FROM patients")
h3.registerTempTable("hos3")
h4 = sqlContext.sql("SELECT * FROM hos3, hos WHERE
hos3.patientid=hos.patientid")
def parsePoint(x):
return LabeledPoint(x[2], x[4:])
parsedData = g4.map(parsePoint)
# Build the model
model = LinearRegressionWithSGD.train(parsedData)
# Evaluate the model on training data
valuesAndPreds = parsedData.map(lambda p:
(p.label, model.predict(p.features)))
MSE = valuesAndPreds.map(lambda
```

```
(v, p): (v - p)**2).reduce(lambda x, y: x + y) / valuesAndPreds.count()
print("Mean Squared Error = " + str(MSE))
```

Spark implementation for computing correlation between the expression levels

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row
from pyspark.mllib.stat import Statistics
from pyspark.mllib.linalg import Vectors
#sc = SparkContext(appName="App")
sqlContext = SQLContext(sc)
status = sc.textFile('/home/ubuntu/HospitalMetaData-10-10.txt')
header = status.first() #extract header
hospital= status.filter(lambda x:x !=header)
sparts = status.map(lambda l: l.split(", "))
hospitalframe = sparts.map(lambda p: Row(statusid=int(p[0]),
target=int(p[1]), position = long(p[2]),
length=int(p[3]), function=int(p[4])))
schema = sqlContext.createDataFrame(statusframe)
schemaStatus.registerTempTable("status")
patients = sc.textFile('/home/ubuntu/PatientMetaData-10-10.txt')
header = patients.first() #extract header
patients = patients.filter(lambda x:x !=header)
pparts = patients.map(lambda l: l.split(", "))
patientsframe = pparts.map(lambda p: Row(patientid=int(p[0]),
age=int(p[1]), gender=int(p[2]), zipcode=int(p[3]),
disease=int(p[4]), patientStatus = float(p[5])))
schemaPatients = sqlContext.createDataFrame(patientsframe)
schemaPatients.registerTempTable("patients")
geo = sc.textFile('/home/ubuntu/GEO-10-10.txt')
header = geo.first() #extract header
geo = geo.filter(lambda x:x !=header)
geoparts = geo.map(lambda l: l.split(", "))
geoframe = geoparts.map(lambda p: Row(statusid=int(p[0]),
patientid=int(p[1]), patientstatus = float(p[2])))
schemaGEO = sqlContext.createDataFrame(geoframe)
schemaGEO.registerTempTable("geo")
g = sqlContext.sql("SELECT p.patientid, p.disease,
e.hospitalid, e.FROM patients AS p,
geo AS e WHERE p.dise=18
AND p.patientid = e.patientid")
g1=g.groupBy('patientid').pivot('hospitalid').sum('exValue')
def parseFunc(x):
return Vectors.dense(x[1:])
parsedData = g1.map(parseFunc)
pearsonCorr = Statistics.corr(parsedData)
print(str(pearsonCorr).replace('nan', 'NaN'))
```

2. **Apa itu Analisis Deskriptif?**

Analitik deskriptif melihat data secara statistik untuk memberi tahu kita apa yang terjadi di masa lalu. Analitik deskriptif membantu bisnis memahami bagaimana kinerjanya dengan memberikan konteks untuk membantu pemangku kepentingan menafsirkan informasi. Ini bisa dalam bentuk visualisasi data seperti grafik, grafik, laporan, dan dasbor.

Bagaimana analitik deskriptif dapat membantu di dunia nyata? Dalam pengaturan layanan kesehatan, misalnya, katakan bahwa jumlah orang yang luar biasa tinggi dirawat di ruang gawat darurat dalam waktu singkat. Analitik deskriptif memberi tahu kita bahwa ini terjadi dan menyediakan data waktu-nyata dengan semua statistik yang sesuai (tanggal kejadian, volume, detail pasien, dll.).

Apa itu Analisis Diagnostik?

Analitik diagnostik mengambil data deskriptif selangkah lebih maju dan memberikan analisis yang lebih mendalam untuk menjawab pertanyaan: Mengapa ini terjadi? Seringkali, analisis diagnostik disebut sebagai analisis akar penyebab. Ini termasuk menggunakan proses seperti penemuan data, penambangan data, dan menelusuri dan menelusuri.

Dalam contoh layanan kesehatan yang disebutkan sebelumnya, analitik diagnostik akan mengeksplorasi data dan membuat korelasi. Misalnya, ini dapat membantu Anda menentukan bahwa semua gejala pasien covid 19 — demam tinggi, batuk kering, dan kelelahan — menunjuk

ke agen infeksi yang sama. Anda sekarang memiliki penjelasan untuk lonjakan volume tiba-tiba di UGD.

Apa itu Predictive Analytics?

Analitik prediktif mengambil data historis dan memasukkannya ke dalam model pembelajaran mesin yang mempertimbangkan tren dan pola utama. Model ini kemudian diterapkan pada data saat ini untuk memprediksi apa yang akan terjadi selanjutnya.Kembali ke contoh rumah sakit kami, analitik prediktif dapat memperkirakan lonjakan pasien yang dirawat di UGD dalam beberapa minggu ke depan. Berdasarkan pola dalam data, penyakit ini menyebar dengan sangat cepat.

Apa itu Analisis Preskriptif?

Analitik preskriptif membawa data prediktif ke tingkat berikutnya. Sekarang setelah Anda memiliki gagasan tentang apa yang akan terjadi di masa depan, apa yang harus kita lakukan? Ini menyarankan berbagai tindakan dan menguraikan apa implikasi potensial bagi masing-masing. Kembali ke contoh rumah sakit kami: sekarang setelah Anda tahu penyakitnya menyebar, alat analisis preskriptif dapat menyarankan agar kita menambah jumlah staf untuk menangani secara memadai masuknya pasien.

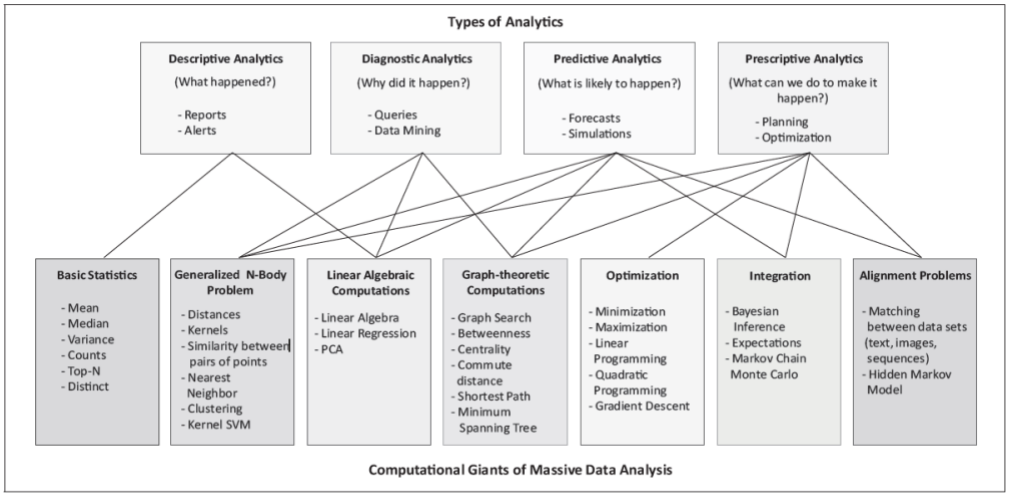


Figure 1.1: Mapping between types of analytics and computational tasks or 'giants'

3. Big Data dalam bidang Kesehatan

Menurut pendapat saya,Ekosistem kesehatan terdiri dari berbagai entitas termasuk penyedia layanan kesehatan(dokter perawatan primer, spesialis, atau rumah sakit), pembayar (pemerintah, kesehatan swasta,perusahaan asuransi, pengusaha), farmasi, perangkat dan layanan medis perusahaan, IT solusi dan layanan perusahaan, dan pasien. Proses penyediaan layanan kesehatan melibatkan data kesehatan besar yang ada dalam bentuk yang berbeda (terstruktur atau tidak terstruktur), disimpan dalamsumber data yang berbeda (seperti database relasional, atau server file) dan dalam berbagai Format. Untuk mempromosikan lebih banyak koordinasi perawatan di beberapa provider yang terlibat dengan pasien, informasi klinis mereka semakin diagregasi dari berbagai sumber ke sistem catatan kesehatan elektronik (EHR). EHRs menangkap dan menyimpan informasi kesehatan pasien dan tindakan penyedia termasuk hasil laboratorium tingkat individu, diagnostik, dan data demografis. Meskipun penggunaan utama EHRs adalah untuk menjaga semua medis data untuk pasien individu dan untuk memberikan akses efisien ke data yang disimpan pada titik perawatan, EHRs dapat menjadi sumber untuk informasi agregat berharga tentang pasien secara keseluruhan populasi .

Dengan ledakan data klinis saat ini data masalah bagaimana untuk mengumpulkan data dari didistribusikan dan heterogen kesehatan sistem TI dan bagaimana menganalisis skala besar klinis data menjadi sangat penting. Sistem data besar dapat digunakan untuk pengumpulan data dari berbagai pemangku kepentingan (pasien, dokter, pembayar, dokter, spesialis, dll) dan sumber data yang berbeda (database, format terstruktur dan tidak terstruktur, dll). Sistem analitik data besar memungkinkan analitik data klinis berskala besar dan memfasilitasi pengembangan perawatan kesehatan yang lebih efisien aplikasi, meningkatkan akurasi prediksi dan membantu dalam pengambilan keputusan yang tepat waktu. Mari kita lihat beberapa aplikasi kesehatan yang dapat mengambil keuntungan dari sistem Big Data:

_ Epidemiologi surveilans: sistem surveilans epidemiologi studi distribusi dan determinan negara atau peristiwa yang berhubungan dengan kesehatan pada populasi tertentu menerapkan studi ini untuk diagnosis penyakit di bawah pengawasan di tingkat nasional untuk mengontrol masalah kesehatan. Sistem EHR termasuk hasil laboratorium tingkat individu, data diagnostik, perawatan, dan demografis. Kerangka data besar dapat digunakan untuk mengintegrasikan data dari beberapa sistem EHR dan analisis data yang tepat waktu secara efektif dan akurat memperkirakan wabah, pengawasan kesehatan tingkat populasi upaya deteksi penyakit dan pemetaan kesehatan masyarakat.

_ Pasien kesamaan berbasis keputusan intelijen aplikasi: kerangka data besar dapat digunakan untuk menganalisis data EHR untuk mengekstrak sekelompok catatan pasien yang paling mirip

pasien target tertentu. Catatan pasien Clustering juga dapat membantu dalam mengembangkan aplikasi prognosis medis yang memprediksi kemungkinan hasil dari suatu penyakit pasien berdasarkan hasil untuk pasien serupa.

_ Obat yang merugikan Events prediksi: Big data kerangka dapat digunakan untuk menganalisis Data EHR dan memprediksi pasien mana yang paling berisiko karena memiliki respons negatif terhadap obat tertentu berdasarkan reaksi obat yang merugikan pasien lain.

_ Mendeteksi anomali klaim: perusahaan asuransi Heath dapat memanfaatkan data besar Analisis klaim asuransi kesehatan untuk mendeteksi penipuan, penyalahgunaan, limbah, dan Kesalahan.

_ Obat berbasis bukti: sistem data besar dapat menggabungkan dan menganalisis data dari berbagai sumber, termasuk hasil laboratorium tingkat individu, diagnostik, pengobatan dan data demografis, untuk mencocokkan perawatan dengan hasil, memprediksi pasien berisiko penyakit corona. Sistem untuk obat berbasis bukti memungkinkan penyedia untuk membuat keputusan tidak hanya berdasarkan persepsi mereka sendiri tetapi juga dari bukti yang tersedia.

_ Real-time pemantauan Kesehatan: perangkat elektronik Wearable memungkinkan non-invasif dan pemantauan parameter fisiologis secara terus-menerus. Perangkat yang dapat dikenakan ini mungkin dalam berbagai bentuk seperti ikat pinggang dan pergelangan tangan-band. Penyedia layanan kesehatan dapat menganalisis
data kesehatan yang dikumpulkan untuk menentukan kondisi kesehatan atau anomali apa pun. Besar sistem data untuk analisis data real-time dapat digunakan untuk analisis volume besar data yang bergerak cepat dari perangkat yang dapat dikenakan dan perangkat lain di rumah Real-Time pemantauan kesehatan pasien dan prediksi merugikan peristiwa

Analytics flow dari aplikasi Covid-19

Pengumpulan data

Mari kita asumsikan bahwa kita memiliki dataset mentah yang tersedia baik dalam database SQL atau sebagai mentah file teks. Untuk mengimpor dataset dari database SQL ke dalam tumpukan data besar, kita dapat menggunakan Konektor SQL. Sedangkan untuk mengimpor file dataset mentah, konektor sumber-Sink dapat berguna.

Persiapan data

Dalam langkah persiapan data, kita mungkin harus melakukan pembersihan data (untuk menghapus Catatan korup) dan perselisihan data (untuk mengubah catatan dalam format yang berbeda format yang konsisten).

Jenis analisis

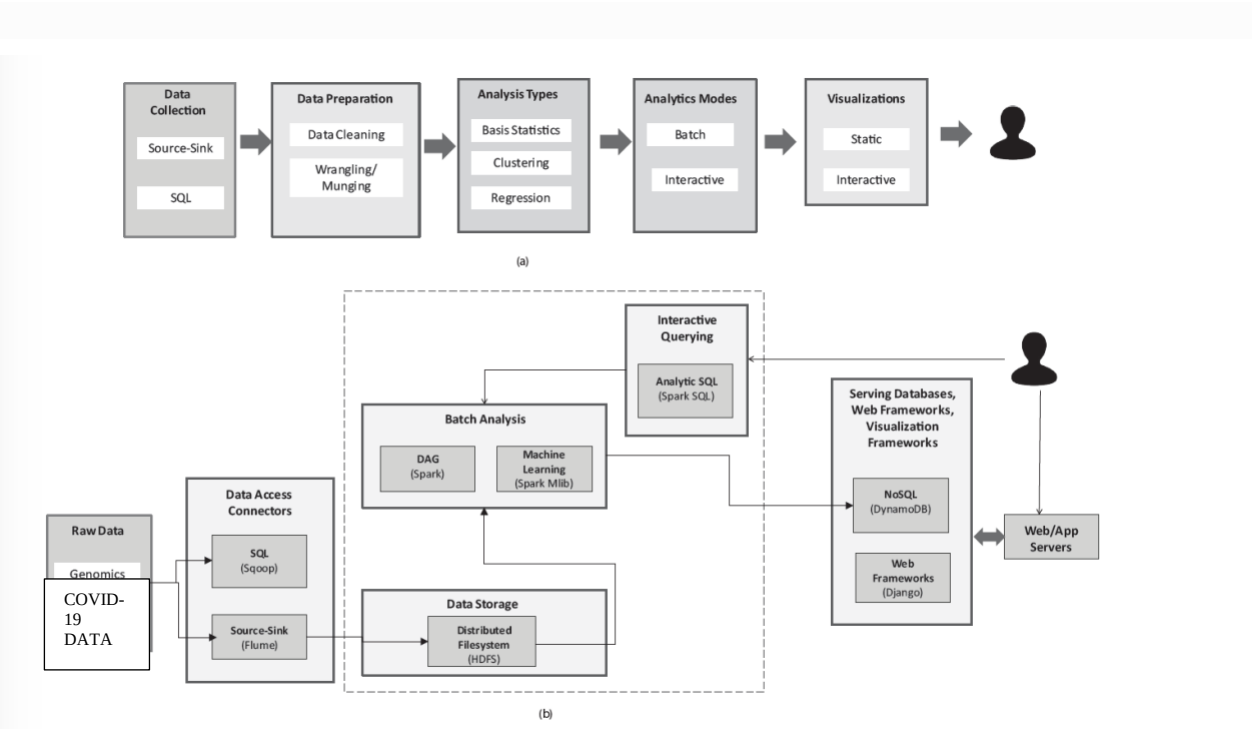
Mari kita katakan, untuk aplikasi ini kita ingin melakukan dua jenis analisis sebagai berikut:
(1) memprediksi pasien suspect covid-19 yang dicurigai dalam pergerakan sebelum dikarantina .
(2) menemukan korelasi antara nilai ekspresi dari semua pasangan hospitality untuk menemukan status pasien RS .yang memiliki pola ekspresi yang sama dan status pasien RS yang memiliki pola ekspresi menentang. Analisis pertama datang di bawah
Analisis regresi kategori, di mana model regresi dapat dibangun untuk memprediksi status pasien Respon. Variabel target untuk model regresi adalah memprediksi pasien suspect covid-19 yang dicurigai dalam pergerakan sebelum dikarantina dan variabel independen adalah nilai ekspresi status pasien RS covid19 . Jenis analisis kedua datang di bawah kategori statistik dasar, di mana kita menghitung korelasi antara ekspresi nilai dari semua pasangan status pasien RS covid19 .

Mode analisis

Berdasarkan jenis analisis ditentukan langkah sebelumnya, kita tahu bahwa analisis mode diperlukan untuk aplikasi akan batch dan interaktif.

Visualisasi
Aplikasi Front end untuk memvisualisasikan hasil analisis akan menjadi dinamis dan interaktif.

Mapping Analytics Flow untuk Big Data Stack
Dengan alur analisis untuk aplikasi yang dibuat, sekarang kita dapat memetakan pilihan di setiap langkah aliran ke tumpukan data besar.
Sebelum kita dapat membangun model regresi, kita harus melakukan beberapa transformasi dan bergabung untuk membuat data yang cocok untuk membangun model.
Saya memilih dengan seperangkat fungsi tertentu dan bergabung dengan gen meta-data dengan pasien data meta-data dan microarray. Selanjutnya, kita menpivot hasil untuk mendapatkan nilai ekspresi untuk Setiap jenis status pasien RS covid19 untuk setiap pasien. Kemudian kita pilih pasien-ID,person status dan placetype dari Meta-data pasien. Selanjutnya, kami bergabung dengan tabel yang diperoleh dalam dua langkah sebelumnya untuk menghasilkan tabel baru yang memiliki semua data dalam format yang tepat untuk membangun model regresi.
Kami memilih pasien dengan status tertentu dan bergabung dengan hasil dengan tabel microarray. Selanjutnya, kita Pivot tabel di langkah sebelumnya untuk mendapatkan nilai ekspresi untuk semua untuk setiap pasien. Kami menggunakan tabel ini untuk membuat korelasi matriks memiliki korelasi antara nilai ekspresi semua pasangan.



a) Analytics flow for covid-19 data analysis
b) Using big data stack for analysis of covid-19 data

4. Big Data Analytics Framework
Berikut ini saya akan menjelaskan mengenai big data analytics framwork
1. Spark Mllib
Spark Mllib adalah Perpustakaan pembelajaran mesin Spark yang menyediakan implementasi

berbagai algoritma pembelajaran mesin termasuk klasifikasi, regresi, Clustering, penyaringan kolaboratif dan pengurangan dimensionalitas. Mllib api dibangun di atas kumpulan data tangguh yang didistribusikan oleh Spark (RDDs). Mllib juga menyediakan jenis data tingkat tinggi seperti Vector, Labeledpoint, rating and Matrix, yang didukung oleh rdds.

Manfaat menggunakan Mllib melalui Perpustakaan pembelajaran mesin adalah bahwa ia menyediakan implementasi paralel algoritma pembelajaran mesin dan dapat memproses dataset terdistribusi yang besar. Spark Mllib menyediakan api untuk Python, Scala, dan bahasa pemrograman Java. Gambar ini menunjukkan berbagai komponen Spark Mllib.

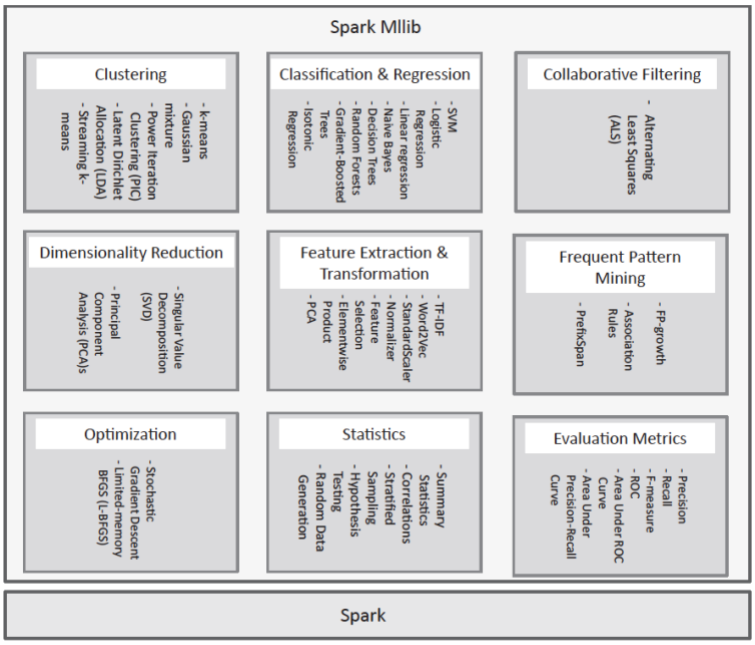


Figure 11.1: Spark MLlib components

2. H2O

H2O adalah open source kerangka analisis prediktif yang menyediakan implementasi dari berbagai algoritma pembelajaran mesin untuk Clustering, klasifikasi, dan dimensionalitas Pengurangan. H2O menyediakan api untuk bahasa pemrograman Python, Scala, R dan Java. H2O juga menyediakan gaya notebook antarmuka web disebut H2O yang memungkinkan pengguna untuk mengimpor data dari berbagai sumber, membangun model pembelajaran mesin dan membuat prediksi menggunakan model. H2O dapat dijalankan sebagai kluster berdiri sendiri atau di atas kluster hadoop atau Spark yang ada. H2O's air soda Perpustakaan mengintegrasikan mesin Machine Learning H2O dengan Spark. H2O dapat terhubung ke berbagai sumber data seperti HDFS, S3, SQL, dan NoSQL.

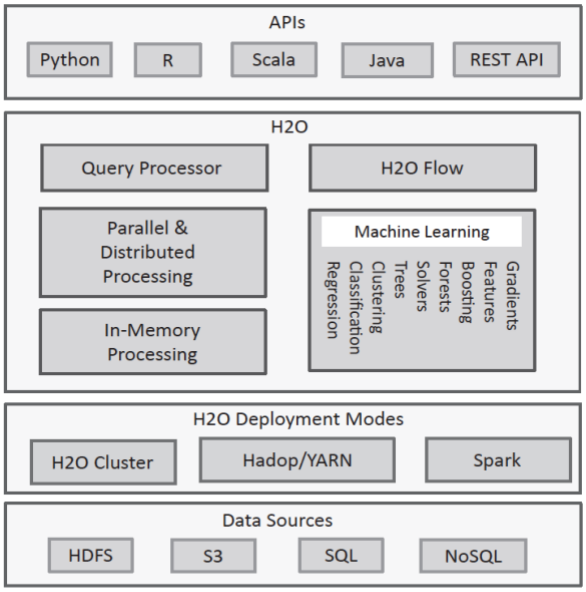
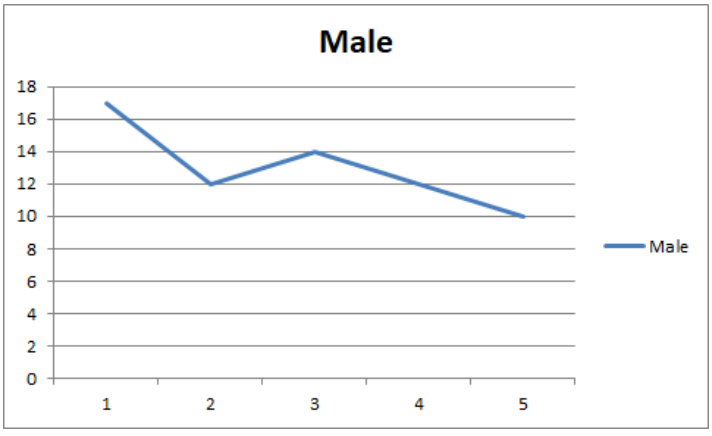


Figure 11.2: H2O components

5. **Grafik Garis atau Bagan**

Grafik garis diwakili oleh sekelompok titik data yang bergabung bersama oleh garis lurus. Masing-masing titik data ini menggambarkan hubungan antara sumbu horizontal dan vertikal pada grafik.

grafik garis atau grafik garis



Saat membuat bagan garis, Anda dapat memutuskan untuk memasukkan poin data atau tidak.

Jenis Grafik Garis

Grafik Garis Sederhana

Dalam grafik garis sederhana, hanya satu garis diplot pada grafik. Salah satu sumbu mendefinisikan variabel independen sedangkan sumbu lainnya berisi variabel dependen.

Grafik Garis Berganda

Beberapa grafik garis berisi dua atau lebih garis yang mewakili lebih dari satu variabel dalam suatu dataset. Jenis grafik ini dapat digunakan untuk mempelajari dua atau lebih variabel selama periode waktu yang sama.

Grafik Garis Kompon

Grafik garis majemuk adalah perpanjangan dari grafik garis sederhana, yang digunakan ketika berhadapan dengan kelompok data yang berbeda dari dataset yang lebih besar. Setiap garis dalam grafik garis majemuk diarsir ke bawah ke sumbu x.

Dalam grafik garis majemuk, setiap kelompok data yang diwakili oleh grafik garis sederhana ditumpuk satu sama lain.

Penggunaan Grafik Garis

- Ini membantu dalam mempelajari tren data selama periode waktu tertentu.
- Mereka mudah dibaca dan plot.

Bar Charts

Bagan batang adalah untuk konsep perbandingan dan persentase di antara faktor atau set data. Pengguna dapat menetapkan berbagai pilihan berbeda untuk responden Anda, misalnya, data pasien bulanan atau tahunan Anda dapat melihat bagan batang mirip dengan bagan kolom apa yang terletak pada sumbu X-nya.

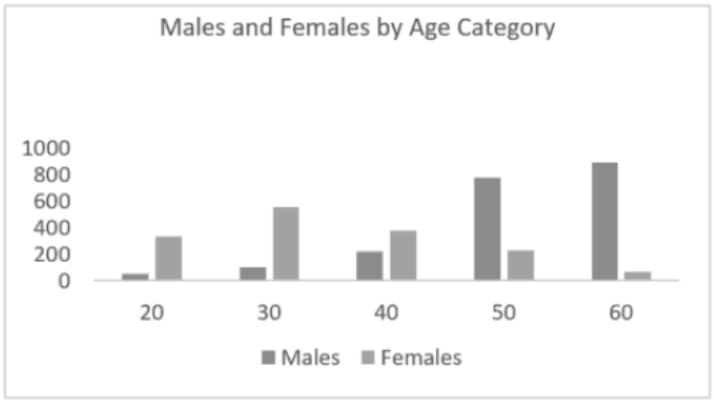


Figure 4.6 Column (bar) chart showing the distribution of two gender variables by age.

Jika kita tidak yakin kapan harus memasukkan bagan batang untuk pekerjaan kita, kita harus memikirkan jenis spesifik dari data asli Anda dan preferensi pribadi Anda. Biasanya, dibandingkan dengan jenis grafik lainnya, diagram batang lebih baik untuk menunjukkan dan membandingkan set data atau angka yang sangat besar.

Jenis-jenis Grafik Batang

Bagan Batang Dikelompokkan

Bagan batang yang dikelompokkan digunakan ketika kumpulan data memiliki subkelompok yang perlu divisualisasikan pada grafik. Setiap subkelompok biasanya dibedakan dari yang lain dengan memberi naungan dengan warna yang berbeda.

Bagan Batang Tertumpuk

Grafik batang yang ditumpuk juga digunakan untuk menampilkan subkelompok dalam dataset. Tetapi dalam kasus ini, bar persegi panjang yang mendefinisikan masing-masing kelompok ditumpuk di atas satu sama lain.

Bagan Batang Tersegmentasi

Ini adalah jenis bagan batang bertumpuk di mana setiap batang bertumpuk menunjukkan persentase dari nilai diskritnya dari nilai total. Persentase total adalah 100%

Keuntungan dari Bar Chart

- Meringkas sejumlah besar data dalam bentuk yang dapat dimengerti.
- Mudah diakses oleh khalayak luas.

Kerugian dari Bar Chart

- Itu tidak mengungkapkan asumsi utama seperti sebab, efek, pola, dll.
- Mungkin memerlukan penjelasan lebih lanjut.

Pie Charts

Pie chart baik untuk menggambarkan dan menunjukkan sampel memecah dalam dimensi individu. Ini dalam bentuk kue untuk menunjukkan hubungan antara utama dan sub-kategori data Anda. Ini baik untuk digunakan ketika Anda berurusan dengan kelompok data yang dikategorikan, atau jika Anda ingin menunjukkan perbedaan di antara data berdasarkan satu variabel.

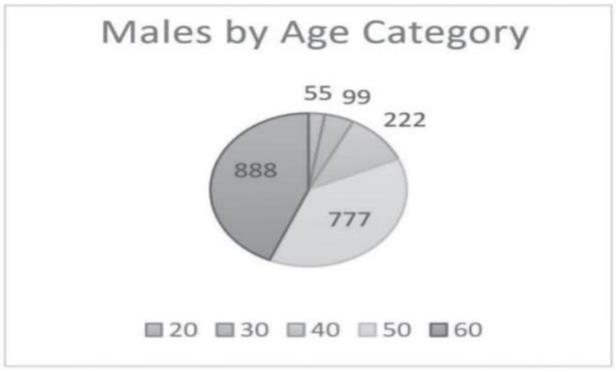


Figure 4.5 Pie chart showing the distribution of one variable (age category) for a sample of males.

Pada kenyataannya, kita dapat memecah kelompok data sampel apa pun ke dalam kategori yang berbeda, misalnya, berdasarkan jenis kelamin atau dalam kelompok usia yang berbeda. Untuk kasus covid-19 kita dapat menggunakan diagram lingkaran untuk mewakili pentingnya satu faktor spesifik pada faktor lainnya. Namun, untuk menganalisis beberapa set data yang berbeda Anda harus pergi untuk bagan kolom. Di dalam pie chart kita bisa dengan mudah mengetahui usia berapa saja yang rentang terinfeksi virus covid-19.

Jenis-jenis Pie Chart

Pie Chart Sederhana

Ini adalah jenis pai bagan yang paling dasar dan juga bisa disebut pai bagan.

Pie Chart yang meledak

Dalam bagan pie yang meledak, salah satu sektor lingkaran dipisahkan (atau meledak) dari bagan. Ini digunakan untuk memberikan penekanan pada elemen tertentu dalam kumpulan data.

Pie of Pie

Seperti namanya, pai pie adalah bagan yang menghasilkan bagan pai yang sama sekali baru (biasanya kecil) dari yang sudah ada. Ini dapat digunakan untuk mengurangi kekacauan dan menekankan pada kelompok elemen tertentu.

Bar Pie

Ini mirip dengan pai pie, dengan perbedaan utama adalah bahwa diagram batang adalah apa yang dihasilkan dalam kasus ini daripada diagram lingkaran.

Bagan Pie 3D

Ini adalah jenis diagram lingkaran yang direpresentasikan dalam ruang 3 dimensi.

Penggunaan Pie Chart

Ini meringkas data menjadi bentuk yang menarik secara visual.

Ini cukup sederhana dibandingkan dengan banyak tipe grafik.

HISTOGRAM

Bagan Histogram

Grafik histogram memvisualisasikan frekuensi data diskrit dan kontinu dalam dataset menggunakan bar persegi panjang yang digabungkan. Setiap bar persegi panjang mendefinisikan jumlah elemen yang jatuh ke dalam interval kelas yang telah ditentukan.

Jenis-jenis Grafik Histogram

Bagan histogram diklasifikasikan menjadi beberapa bagian tergantung pada distribusinya

• **Distribusi normal**

Bagan histogram yang terdistribusi normal biasanya berbentuk lonceng. Seperti namanya, distribusi ini normal dan merupakan standar untuk bagaimana grafik histogram normal seharusnya terlihat.

• **Distribusi Bimodal**

Dalam bagan histogram terdistribusi secara bimodal, kami memiliki dua kelompok grafik histogram yang berdistribusi normal. Ini dibentuk sebagai hasil dari menggabungkan dua proses dalam sebuah dataset.

• Distribusi yang Timpang

Ini adalah grafik asimetris dengan pick off-center biasanya cenderung pada akhir grafik. Bagan

histogram dapat dikatakan miring kanan atau kiri tergantung pada arah kemana arah puncak

cenderung.

• Distribusi Acak

Jenis bagan histogram ini tidak memiliki pola biasa. Ini menghasilkan beberapa puncak dan juga

bisa disebut distribusi multimodal.

• Distribusi Puncak Puncak

Distribusi ini memiliki struktur yang mirip dengan distribusi normal dengan puncak besar di salah satu ujungnya menjadi faktor pembeda.

• Distribusi Sisir

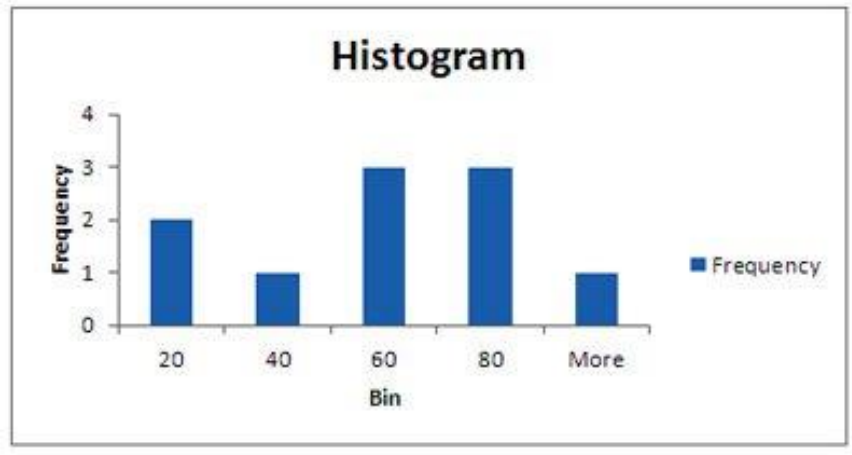
Distribusi sisir memiliki struktur "seperti sisir", di mana batang persegi panjang bergantian antara tinggi dan pendek.

Penggunaan Bagan Histogram

• Ini membantu dalam memvisualisasikan sejumlah besar data.

• Mengungkap variasi, pemusatan, dan distribusi data.

Cons Itu tidak memvisualisasikan nilai yang tepat dalam suatu dataset. Ini hanya memvisualisasikan data kontinu.



SCARTTER CHART

Scatter chart ideal untuk menganalisis bagaimana tujuan yang berbeda diselesaikan di sekitar topik utama dan berbagai dimensinya beberapa saat. Misalnya, kita dapat dengan cepat melihat angka pengurangan ataupun pertambahan kasus covid-19 di Indonesia. Scatter chart memiliki beberapa elemen berbeda: marker, poin, dan garis lurus. Semua faktor ini dapat menunjukkan

danmenghubungkan unit data yang berbeda.

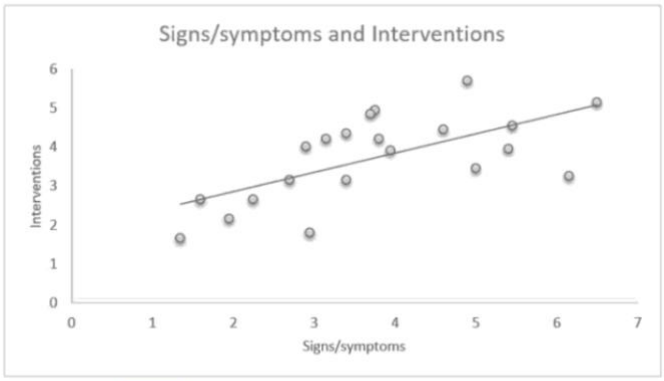


Figure 4.13. Scatter plot with trend line depicting relationships between signs/symptoms (x-axis) and interventions (y-axis)

Kita dapat memilih untuk menggambar bagan sebar hanya di spidol atau garis. Secara umum, marker ideal untuk titik data kecil, sementara garis berguna untuk titik data ukuran besar.

Scatter chart memiliki titik yang sama dengan diagram garis karena keduanya menggunakan sumbu vertikal dan horizontal untuk menunjukkan titik data yang berbeda, tetapi tipe pencar juga dapat menunjukkan tingkat perbedaan dalam satu variabel dengan yang lain, yang dikenal sebagai korelasi. Korelasi bisa positif, negatif, atau sama dengan nol. Yang positif, misalnya, berarti data meningkat secara simultan sebagian besar waktu berdasarkan waktu yang diberikan.