LAPORAN TUGAS BESAR STATISTIKA SAINS DATA (SSD)

Mochammad Aditya Putra Suhendar
120450058 - RB

BACKGROUND

Dalam tugas besar mata kuliah Statitstika Sains Data, dataset yang digunakan adalah dataset covid_uts.csv dimana dataset tersebut sudah pernah dipakai untuk menjadi dataset dalam Ujiang Tengah Semester (UTS).

Terdapat 4 bab pembahasan dalam laporan tugas besar mata kuliah Statistika Sains Data (SSD):

1. Data Wrangling
2. Data Visualization
3. Data Processing
4. Implementation Model

DATA WRANGLING

Data wrangling adalah proses pembersihan, penataan, dan pengayaan data mentah ke dalam format yang diinginkan untuk menghasilkan pengambilan keputusan yang lebih baik dalam waktu yang lebih singkat. Berikut data yang saya tampilkan menggunakan query:

```
# Menampilkan 5 data teratas
df.head()
```

	Negara	Tanggal	Varian	N_Positif
0	Egypt	2020-05-11	Alpha	0
1	Egypt	2020-05-11	Beta	0
2	Egypt	2020-05-11	Gamma	0
3	Egypt	2020-05-11	Mu	0
4	Egypt	2020-05-11	Omicron	0

head() digunakan untuk menampilkan data teratas pada dataframe. Secara default akan menampilkan 5 data teratas kecuali menginput angka *n* didalam tanda kurung yang ada maka akan menampilkan sejumlah *n* dari data teratas.

```
# Menampilkan 5 data terbawah
df.tail()
```

	Negara	Tanggal	Varian	N_Positif
1485	Malaysia	2021-12-27	Alpha	0
1486	Malaysia	2021-12-27	Beta	0
1487	Malaysia	2021-12-27	Gamma	0
1488	Malaysia	2021-12-27	Mu	0
1489	Malaysia	2021-12-27	Omicron	5

tail() digunakan untuk menampilkan data terbawah pada dataframe. Secara default akan menampilkan 5 data terbawah kecuali menginput angka *n* didalam tanda kurung yang ada maka akan menampilkan sejumlah *n* dari data terbawah.

```
# Informasi data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1490 entries, 0 to 1489
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Negara     1490 non-null   object
 1   Tanggal    1490 non-null   object
 2   Varian     1490 non-null   object
 3   N_Positif  1490 non-null   int64 
dtypes: int64(1), object(3)
memory usage: 46.7+ KB
```

info() digunakan untuk menampilkan informasi detail dataframe, seperti jumlah baris data, nama kolom, jumlah data dan tipe datanya, dsb.

```
# Statistik data
df.describe()
```

	N_Positif
count	1490.000000
mean	133.230872
std	1067.331772
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	18317.000000

describe() digunakan untuk menampilkan deskripsi statistik data. Hanya kolom yang bertipe numerik yang akan ditampilkan statistiknya seperti count, mean, standard deviation, nilai minimum-maximum.

```
# Menampilkan jumlah negara
negaras = df['Negara'].unique()
print('Jumlah Negara: ', len(negaras))
for neg in negaras:
    print('-',neg)
```

```
Jumlah Negara:  7
- Egypt
- Finland
- Germany
- Indonesia
- Italy
- Japan
- Malaysia
```

```
# Menampilkan jumlah varian COVID
varians = df['Varian'].unique()
print('Jumlah Varian Covid: ' , len(varians))
for v in varians:
    print('-',v)
```

```
Jumlah Varian Covid:  5
- Alpha
- Beta
- Gamma
- Mu
- Omicron
```

unique() digunakan untuk menampilkan nilai unik dari suatu kolom didalam sebuah dataframe.

```
# Mengubah kolom menjadi baris
pd.melt(df)
```

	variable	value
0	Negara	Egypt
1	Negara	Egypt
2	Negara	Egypt
3	Negara	Egypt
4	Negara	Egypt
...	...	...
5955	N_Positif	0
5956	N_Positif	0
5957	N_Positif	0
5958	N_Positif	0
5959	N_Positif	5

5960 rows × 2 columns

melt() digunakan untuk mengubah kolom yang ada menjadi baris.

drop() digunakan untuk menghapus kolom dalam sebuah dataframe.

```
# Menghapus kolom tanggal
df.drop(columns=['Tanggal'])
```

	Negara	Varian	N_Positif
0	Egypt	Alpha	0
1	Egypt	Beta	0
2	Egypt	Gamma	0
3	Egypt	Mu	0
4	Egypt	Omicron	0
...	...	...	...
1485	Malaysia	Alpha	0
1486	Malaysia	Beta	0
1487	Malaysia	Gamma	0
1488	Malaysia	Mu	0
1489	Malaysia	Omicron	5

1490 rows × 3 columns

```
# Mengganti nama kolom
df.rename(columns={'Varian': 'Variant'}, inplace=True)
df
```

	Negara	Tanggal	Variant	N_Positif
0	Egypt	2020-05-11	Alpha	0
1	Egypt	2020-05-11	Beta	0
2	Egypt	2020-05-11	Gamma	0
3	Egypt	2020-05-11	Mu	0
4	Egypt	2020-05-11	Omicron	0
...	...	...	...	...
1485	Malaysia	2021-12-27	Alpha	0
1486	Malaysia	2021-12-27	Beta	0
1487	Malaysia	2021-12-27	Gamma	0
1488	Malaysia	2021-12-27	Mu	0
1489	Malaysia	2021-12-27	Omicron	5

1490 rows × 4 columns

rename() digunakan untuk mengganti nama dari kolom yang ada didalam sebuah dataframe.

```
# Mencari negara dengan data terbesar hingga terkecil
df.sort_values('N_Positif', ascending=False)
```

	Negara	Tanggal	Variant	N_Positif
530	Germany	2021-05-03	Alpha	18317
525	Germany	2021-04-19	Alpha	18098
520	Germany	2021-04-05	Alpha	14485
535	Germany	2021-05-17	Alpha	14430
515	Germany	2021-03-22	Alpha	11445
...	...	...	...	...
688	Indonesia	2020-11-09	Mu	0
687	Indonesia	2020-11-09	Gamma	0
686	Indonesia	2020-11-09	Beta	0
685	Indonesia	2020-11-09	Alpha	0
0	Egypt	2020-05-11	Alpha	0

1490 rows × 4 columns

sort_values() digunakan untuk mengurutkan data terbesar hingga terkecil ataupun sebaliknya, dalam kasus ini saya mengurutkan data jumlah positif di 7 negara dari yang terbesar hingga terkecil.

```
# Grouping negara dengan variant COVID
df2=df.groupby(['Negara', 'Variant']).sum()
df2
```

Negara	Variant	N_Positif
Egypt	Alpha	29
	Beta	0
	Gamma	0
	Mu	0
	Omicron	1
Finland	Alpha	6800
	Beta	1213
	Gamma	19
	Mu	5
	Omicron	0
Germany	Alpha	104138
	Beta	2303
	Gamma	858
	Mu	17
	Omicron	2270
Indonesia	Alpha	81
	Beta	22
	Gamma	0
	Mu	0
	Omicron	130
Italy	Alpha	26877
	Beta	116
	Gamma	2488
	Mu	83
	Omicron	526
Japan	Alpha	49841
	Beta	101
	Gamma	120
	Mu	3
	Omicron	150
Malaysia	Alpha	33
	Beta	273
	Gamma	0
	Mu	0
	Omicron	17

groupby() digunakanuntuk melakukan perhitungan kelompok dan menggabungkan kolom berdasarkan nilai unik sesuai kolom yang dipilih.

```
df3 = df2.query("Variant=='Alpha'| Variant=='Beta' | Variant=='Omicron' ")
df3.reset_index(inplace=True)
df3
```

	Negara	Variant	N_Positif
0	Egypt	Alpha	29
1	Egypt	Beta	0
2	Egypt	Omicron	1
3	Finland	Alpha	6800
4	Finland	Beta	1213
5	Finland	Omicron	0
6	Germany	Alpha	104138
7	Germany	Beta	2303
8	Germany	Omicron	2270
9	Indonesia	Alpha	81
10	Indonesia	Beta	22
11	Indonesia	Omicron	130
12	Italy	Alpha	26877
13	Italy	Beta	116
14	Italy	Omicron	526
15	Japan	Alpha	49841
16	Japan	Beta	101
17	Japan	Omicron	150
18	Malaysia	Alpha	33
19	Malaysia	Beta	273
20	Malaysia	Omicron	17

reset_index() digunakan untuk mereset indeks yang telah terset dan menjadikan indeksnya sebagai default yang berupa bilangan integer.

DATA VISUALIZATION

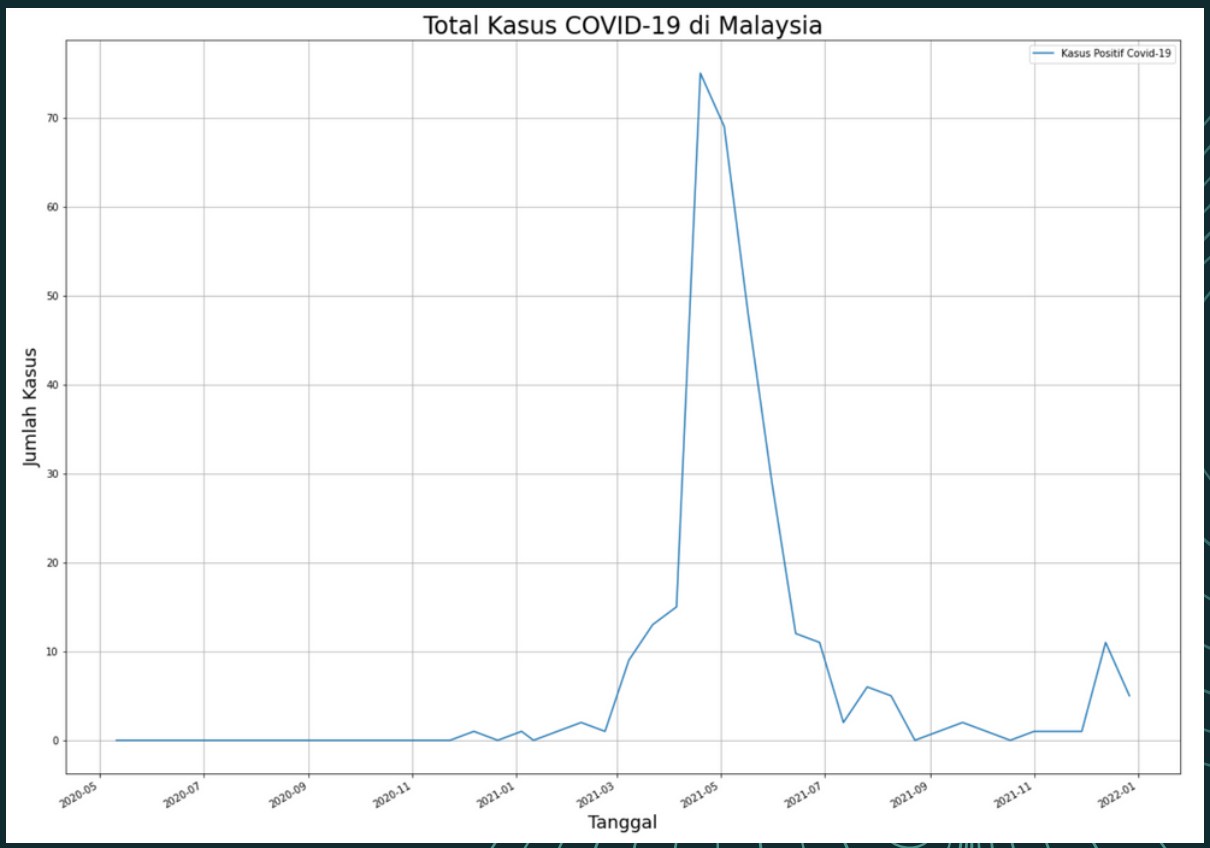
Data Visualization adalah salah satu komunikasi visual modern yang dapat menjadi solusi menyajikan suatu data agar lebih mudah dipahami. Dari segi bahasa bisa diartikan tampilan visual berupa grafis dari informasi dan data tertentu.

Dalam tugas besar ini saya akan memvisualisasikan dalam bentuk data dan plot dengan menggunakan query:

```
df_mal = df.groupby(['Negara','Tanggal']).sum()
df_mal = df_mal.loc[['Malaysia']].reset_index()
df_mal ['Tanggal'] = pd.to_datetime(df_mal['Tanggal'])
df_mal = df_mal.set_index('Tanggal')
```

```
plt.figure(figsize=(20,15))
ax = plt.gca()
df_mal.plot(ax=ax)
ax.grid()
ax.set_xlabel('Tanggal',fontsize=18)
ax.set_ylabel('Jumlah Kasus',fontsize=18)
ax.legend(['Kasus Positif Covid-19'])
ax.set_title('Total Kasus COVID-19 di Malaysia',fontsize=24)
plt.show()
```

Didapatkan hasil plot sebagai berikut:



Dari grafik diatas dapat dilihat bahwa jumlah kasus COVID-19 di Malaysia pada bulan Mei tahun 2020 tercatat 0 kasus, dimulai pada bulan Maret 2021 terjadi lonjakan kasus yang mengakibatkan lonjakan yang sangat tinggi di bulan April 2021. Dilakukan query df_mal untuk mengecek kebenaran dari grafik diatas.

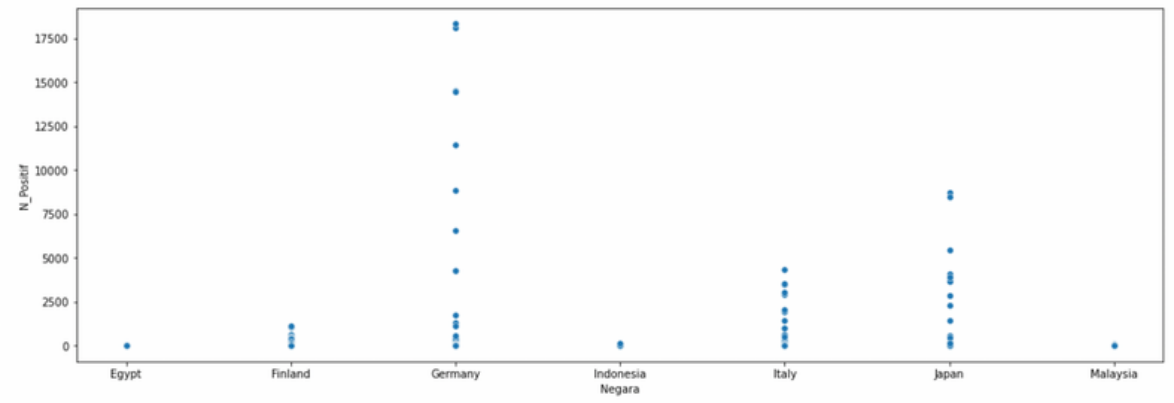
Tanggal	Negara	N_Positif
2020-05-11	Malaysia	0
2020-05-25	Malaysia	0
2020-06-08	Malaysia	0
2020-06-22	Malaysia	0
2020-07-20	Malaysia	0
2020-08-03	Malaysia	0
2020-08-17	Malaysia	0
2020-08-31	Malaysia	0
2020-09-14	Malaysia	0
2020-10-12	Malaysia	0
2020-10-26	Malaysia	0
2020-11-09	Malaysia	0
2020-11-23	Malaysia	0
2020-12-07	Malaysia	1
2020-12-21	Malaysia	0
2021-01-04	Malaysia	1
2021-01-11	Malaysia	0
2021-01-25	Malaysia	1
2021-02-08	Malaysia	2
2021-02-22	Malaysia	1
2021-03-08	Malaysia	9
2021-03-22	Malaysia	13
2021-04-05	Malaysia	15
2021-04-19	Malaysia	75
2021-05-03	Malaysia	69
2021-05-17	Malaysia	18
2021-05-31	Malaysia	29
2021-06-14	Malaysia	12
2021-06-28	Malaysia	11
2021-07-12	Malaysia	2
2021-07-26	Malaysia	6
2021-08-09	Malaysia	5
2021-08-23	Malaysia	0
2021-09-06	Malaysia	1
2021-09-20	Malaysia	2
2021-10-04	Malaysia	1
2021-10-18	Malaysia	0
2021-11-01	Malaysia	1
2021-11-15	Malaysia	1
2021-11-29	Malaysia	1
2021-12-13	Malaysia	11
2021-12-27	Malaysia	5

```
df_mal
```

Dari hasil data diatas terbukti bahwa data kasus covid di Malaysia terbesar berada pada bulan April 2021 yaitu sebanyak 75 kasus, dimana terjadi lonjakan yang signifikan sejak bulan Maret 2021 dan penurunan yang signifikan hingga bulan Juli 2021 semenjak mencapai titik tertinggi pada bulan April 2021.

Selanjutnya ditampilkan diagram scatterplot dari N_Positif setiap negara, dengan menggunakan query dengan hasil sebagai berikut:

```
sns.scatterplot(data=df, x=df['Negara'], y=df['N_Positif']);
```



Didapatkan diagram secara visual bahwa jumlah kasus Covid-19 terbesar adalah di negara Germany dengan lebih dari 17500 kasus Covid-19 kemudian disusul dengan Japan, Italy, Finland, Indonesia, Malaysia, dan yang terendah adalah Egypt. Kemudian dilakukan data processing guna mengetahui kebenaran dari diagram diatas.

Data Processing

Data Processing merupakan proses pengumpulan data dan dikonversi menjadi sebuah informasi yang bermanfaat dan dapat digunakan.

Dilakukan pemrosesan data untuk mengetahui jumlah kasus Covid-19 dari setiap Negara dengan Variant Covid-19 yaitu Alpha, Beta dan Omicron, berikut query yang menghasilkan data dibawah ini:

```
processing = df3[['Negara']]
negara = processing['Negara'].unique()

alpha = df3[df3['Variant']=='Alpha']
a= alpha[['N_Positif']].values.flatten()
beta = df3[df3['Variant']=='Beta']
b = beta[['N_Positif']].values.flatten()
omicron = df3[df3['Variant']=='Omicron']
o = omicron[['N_Positif']].values.flatten()

datas = {
'Negara': negara,
'Alpha':a,
'Beta':b,
'Omicron': o,
}
df_new = pd.DataFrame(datas)
df_new
```

	Negara	Alpha	Beta	Omicron
0	Egypt	29	0	1
1	Finland	6800	1213	0
2	Germany	104138	2303	2270
3	Indonesia	81	22	130
4	Italy	26877	116	526
5	Japan	49841	101	150
6	Malaysia	33	273	17

Dari data diatas dapat disimpulkan bahwa jumlah kasus tertinggi Covid-19 secara berurutan adalah Gemany, Japan, Italy, Finland, Indonesia, Malaysia dan Egypt.

Berikutnya dilakukan perhitngan similaritas dengan mendefinisikan similarity sebagai jarak r yang dihitung menggunakan Euclidean Distance Pada Negara Finland, Indonesia, Italy, Japan, dan Malaysia dengan hasil matrix 5 negara sebagai berikut:

```
def euclid(x,y):

    return np.linalg.norm(x-y)
lneg = len(temps)
d_matrix = [ [ round(euclid(temps[i],temps[j]),2 ) for j in range(lneg) ] for i in range(lneg)]
```

```
negara1 = df4[['Negara']].values.flatten()
df5 = pd.DataFrame( d_matrix,columns = negara1, index = negara1 )
df5
```

	Finland	Indonesia	Italy	Japan	Malaysia
Finland	0.00	6824.98	20113.83	43055.62	6832.00
Indonesia	6824.98	0.00	26799.09	49760.07	279.42
Italy	20113.83	26799.09	0.00	22967.08	26849.28
Japan	43055.62	49760.07	22967.08	0.00	49808.47
Malaysia	6832.00	279.42	26849.28	49808.47	0.00

Selanjutnya dicari manakah Negara yang memiliki similaritas degan Malaysia dengan query dan hasilnya sebagai berikut:

```
d1 = df5.loc['Malaysia']
d2 = df5.columns.values
def similarity(d):
    return 1/(d+0.001)
d3 = map( lambda x,y:[ similarity(x) ] + [y] ,d1,d2)
d4 = list(d3)
d4.sort(key=lambda x:x[0],reverse=True)
d4 = d4[1:]
d4
```

```
[[0.0035788290787020305, 'Indonesia'],
 [0.00014637000199502312, 'Finland'],
 [3.724494521845855e-05, 'Italy'],
 [2.0076906195333723e-05, 'Japan']]
```

Didapatkan bahwa negara yang memiliki similaritas dengan malaysia secara bururutan adalah Indonesia, Finland, Italy dan Japan.

Implementation Model

Metode Regresi Logistik

Merupakan metode analisis statistika untuk mendeskripsikan hubungan antara variabel terikat yang memiliki dua kategori atau lebih dengan satu atau lebih peubah bebas berskala kategori.

Langkah awal dalam metode ini saya mendefinisikan x sebagai N_Positif dan y sebagai Negara serta mengambil secara acak y_test sebesar 50% atau 0,5 dengan menggunakan query dan hasil heatmap sebagai berikut:

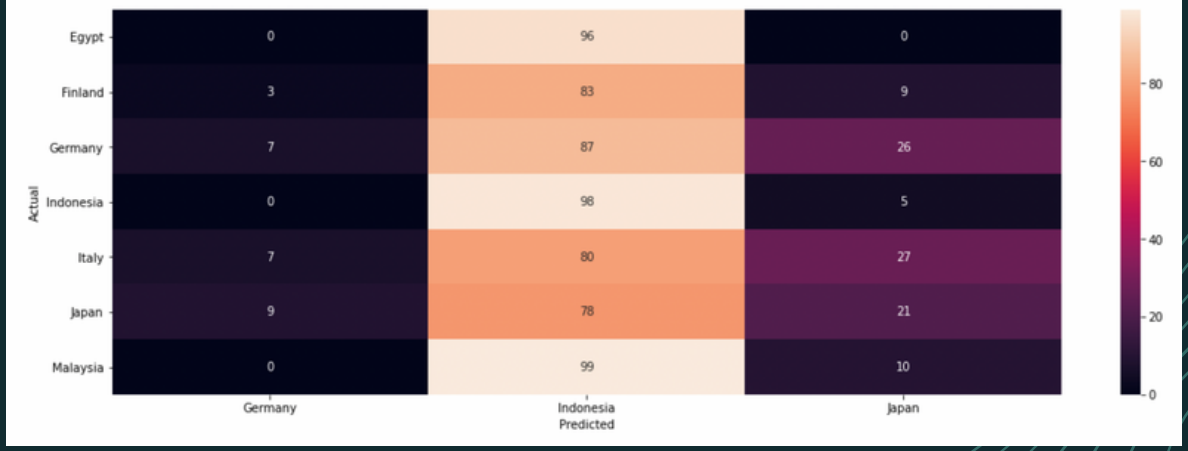
```
X = df[['N_Positif']]
y = df['Negara']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.5, random_state = 0)
```

```
logistic_regression = LogisticRegression(solver='liblinear')
logistic_regression.fit(X_train, y_train)
y_pred = logistic_regression.predict(X_test)
```

```
confusion_matrix = pd.crosstab(y_test, y_pred, rownames=['Actual'], colnames=['Predicted'])
sns.heatmap(confusion_matrix, annot=True)
```

```
<AxesSubplot:xlabel='Predicted', ylabel='Actual'>
```



Kemudian mencari classification_report dengan y_test dan predictions, mencari accuracy_score dengan menggunkan y_test dan predictions dengan menggunakan query dan didapatkan hasil berikut :

```
model=LogisticRegression()
model.fit(X_train,y_train)
predictions=model.predict(X_test)
```

```
print(classification_report(y_test,predictions))
print(accuracy_score(y_test,predictions))
```

	precision	recall	f1-score	support
Egypt	0.00	0.00	0.00	96
Finland	0.00	0.00	0.00	95
Germany	0.30	0.06	0.10	120
Indonesia	0.16	0.95	0.27	103
Italy	0.00	0.00	0.00	114
Japan	0.21	0.19	0.20	108
Malaysia	0.00	0.00	0.00	109
accuracy			0.17	745
macro avg	0.10	0.17	0.08	745
weighted avg	0.10	0.17	0.08	745

0.1691275167785235

Dari hasil diatas terdapat beberapa komponen, seperti accuracy, macro average, weighted average, precision, recall, f1-score dan support dengan hasil negara masing-masing.

Precission merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Dapat disimpulkan dari data diatas bahwa Germany memiliki precision terbesar senilai 0.30 yang berarti Germany merupakan sebuah negara yang terbukti dengan kasus COVID-19 terbesar.

Recall merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Dapat disimpulkan dari data diatas bahwa negara Indonesia memiliki recall terbesar senilai 0,95 dimana prediksi COVID-19 keseluruhan dibandingkan dengan kasus Covid-19 sebenarnya benar bahwa Indonesia memiliki lonjakan yang cukup tinggi di akhir tanggal didalam data tersebut.

F1 Score merupakan perbandingan rata-rata presisi dan recall yang dibobotkan. Dapat disimpulkan dari data diatas bahwa negara Indonesia memiliki F1 score terbesar senilai 0,27. F1 score atau harmonic mean dari precision dan recall. Secara representasi, jika F1-Score punya skor yang baik mengindikasikan bahwa model klasifikasi kita punya precision dan recall yang baik.

Kemudian terdapat accuracy, dimana dalam data tersebut dihasilkan accuracy sebesar 0,1691275167785235 dimana accuracy merupakan rasio prediksi benar dengan keseluruhan data.