



KEMENTERIAN PENDIDIKAN, KEBUDAYAAN,  
RISET, DAN TEKNOLOGI

Puspresnas  
Pusat Prestasi Nasional

BPTI  
Balai Pengembangan  
Talenta Indonesia



77  
PULIH  
LEBIH CEPAT  
BANOKIT  
LEBIH KUAT



G20  
INDONESIA  
2022



Kampus  
Merdeka

INDONESIA JAYA

Universitas Islam Indonesia

SATRIA 20  
DATA 22

# FINAL

# SATRIA DATA 2022

SD20220000074



**SATRIA 20  
DATA 22**

Statistika Ria dan Festival Sains Data  
Universitas Islam Indonesia



**Puspresnas**  
Pusat Prestasi Nasional

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN,  
RISET, DAN TEKNOLOGI

**BPTI**  
Bali Pengembangan  
Talenta Indonesia



77

PULIH  
LEBIH CEPAT  
BANGET  
LEBIH KUAT



G20  
INDONESIA  
2022



Kampus  
Merdeka  
INDONESIA-JAYA

SATRIA 20  
DATA 22  
Universitas Islam Indonesia



# PROBLEM 1

**PREDIKSI STATUS PULANG PASIEN  
FKTP BPJS**

# KONTEN

01

**PEMAHAMAN PERMASALAHAN DAN DATA**

02

**RANCANGAN ANALISIS DAN ALGORITMA**

03

**HASIL ANALISIS, INSIGHT, DAN REKOMENDASI**

04

**KESIMPULAN DAN SARAN**

# 1. PEMAHAMAN PERMASALAHAN DAN DATA

## MASALAH

Pengefektifan anggaran BPJS kesehatan melalui status pulang pasien FKTP BPJS

## TUJUAN

Membuat model klasifikasi dengan skor F1 tertinggi yang dapat memprediksi status pulang pasien FKTP BPJS

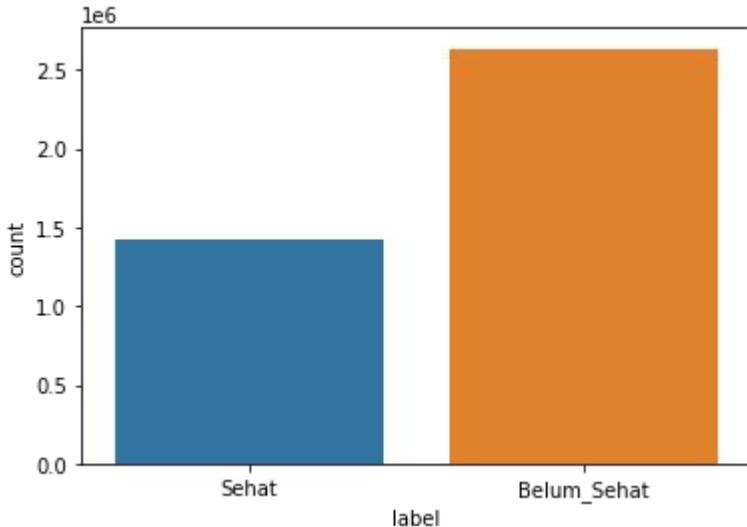
# 1. PEMAHAMAN PERMASALAHAN DAN DATA

## DATASET

- Catatan Kunjungan Pasien FKTP BPJS dari tahun 2015-2020
- Data Latih terdiri dari 4,056,898 baris dan 26 kolom
- **FKP13Class** atau **label** sebagai kolom target dengan 2 kelas Sehat (1) dan Belum\_Sehat (0) dengan rasio 1:1.85
- Data Uji terdiri dari 1,014,225 dan 25 kolom

# 1. PEMAHAMAN PERMASALAHAN DAN DATA

Sebaran Label



# 1. PEMAHAMAN PERMASALAHAN DAN DATA

Nama Kolom

- id\_peserta
- id\_keluarga
- bobot
- id\_kunjungan
- tgl\_datang
- tgl\_pulang
- provinsi\_fktp
- kota\_fktp
- kepemilikan\_fktp
- jenis\_fktp
- tipe\_fktp
- tingkatpel\_fktp
- poli\_fktp
- segmen\_peserta\_fktp
- kode\_nama\_diag
- kode\_diag\_icd
- kode\_diag
- kode\_diag\_fkp15
- provinsi\_fkr
- kota\_fkr
- kepemilikan\_fkr
- jenis\_fkr
- tipe\_fkr
- poli\_fkr
- jenis\_kunjungan\_fktp
- label



# 1. PEMAHAMAN PERMASALAHAN DAN DATA

## **id\_kunjungan dan id\_peserta**

Nilai unik pada feature id\_kunjungan adalah

```
958073632      1
259320919P000811  1
804269476      1
387880120P000493  1
349651019P002103  1
```

..

```
30060219P000030  1
1326480344      1
1121180537      1
319090320P002037  1
1188955927      1
```

Name: id\_kunjungan, Length: 4056898, dtype: int64

Nilai unik pada feature id\_peserta adalah

```
100846584.0    248
60566306.0    235
37301192.0    210
48600054.0    206
73374830.0    196
```

...

```
190005805.0    1
8365752.0     1
12585271.0    1
236919998.0   1
24766757.0    1
```

Name: id\_peserta, Length: 832023, dtype: int64



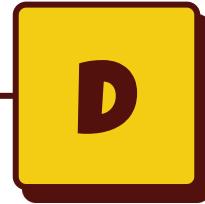
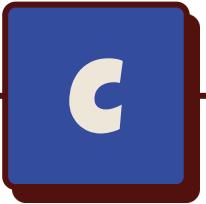
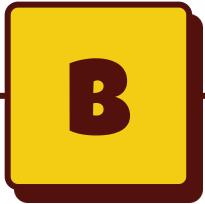
## 2. RANCANGAN ANALISIS DAN ALGORITMA

Diagram Alir

**DATA  
PREPROCESSING**



**MODELLING**



**FEATURE  
ENGINEERING AND  
selection**

**EVALUATION**

## 2. RANCANGAN ANALISIS DAN ALGORITMA

### Data Preprocessing

#### MENGISI NILAI NULL

34% dari total data **kode\_diag\_icd** yang tersedia mempunyai nilai Null. Kemudian diisi “None”

#### UBAH TIPE DATA

Mengubah tipe data pada kolom **tgl\_datang** dan **tgl\_pulang** menjadi Datetime

#### LABEL ENCODING

Pada kolom non-numerik seperti **kode\_diag**, **kode\_diag\_icd**, dan **label**

## 2. RANCANGAN ANALISIS DAN ALGORITMA

*Feature Engineering and Selection*

### KOLOM “LAMA\_RAWAT”

Membuat kolom **lama\_rawat** yang merupakan selisih hari antara tanggal datang dan pulang

### KOLOM “TIDAK\_RUJUK”

Membuat kolom **tidak\_rujuk** untuk melabeli kunjungan yang tidak dirujuk, dilihat dari **jenis\_fkr**

### KOLOM TANGGAL

Mengurai komponen datetime menjadi **hari**, **is\_weekend**, **haripadatahun**, **pekanpadatahun**, **tanggal**, **bulan**, **tahun**, dan **quarter**

## 2. RANCANGAN ANALISIS DAN ALGORITMA

*Feature Engineering and Selection*

### UJI ANOVA

H<sub>0</sub> : variabel numerik x tidak memiliki hubungan dengan variabel y (label)

H<sub>1</sub> : variabel numerik x memiliki hubungan dengan variabel y (label)

### UJI CHI-SQUARE

H<sub>0</sub> : variabel kategorik tidak memiliki hubungan dengan variabel y (label)

H<sub>1</sub> : variabel kategorik memiliki hubungan dengan variabel y (label)



## 2. RANCANGAN ANALISIS DAN ALGORITMA

Uji ANOVA untuk Kolom Numerik

Variabel	F Score	P Value
bobot	131801.917	0.000
tanggal	8390.253	0.000
lama_rawat	7507.107	0.000
hari	4609.151	0.000
quarter	847.399	2.794e-186
bulan	812.241	1.225e-178
pekanpadatahun	530.369	2.388e-177
haripadatahun	444.549	1.120e-98
tahun	13.584	2.280e-04

## 2. RANCANGAN ANALISIS DAN ALGORITMA

Uji Chi-Square untuk Kolom Kategorik

Variabel	Chi Score	P Value
kode_diag	5.541e+09	0.000
kota_fkr	8.147e+07	0.000
kode_fdiag_icd	5.686e+07	0.000
poli_fktp	2.200e+07	0.000
jenis_fkr	2.025e+06	0.000
kepemilikan_fkr	1.793e+06	0.000
tipe_fkr	9.507e+05	0.000
kota_fktp	8.639e+05	0.000
kepemilikan_fktp	8.428e+05	0.000
provinsi_fkr	7.889e+05	0.000

Variabel	Chi Score	P Value
jenis_kunjungan_fktp	6.489e+05	0.000
jenis_fktp	1.915e+05	0.000
segmen_peserta_fktp	1.477e+05	0.000
tipe_fktp	1.337e+05	0.000
tidak_dirujuk	2.112e+04	0.000
provinsi_fktp	8.440e+03	0.000
tingkatpel_fktp	7.424e+02	0.000
poli_fkr	1.952e+02	2.262e-44
is_weekend	1.212e+01	4.980e-04

## 2. RANCANGAN ANALISIS DAN ALGORITMA

*Feature Engineering and Selection*

### KOLOM TERELIMINASI

- `tgl_datang`
- `tgl_pulang`
- `id_kunjungan`
- `id_peserta`
- `id_keluarga`
- `kode_diag_fkp15`

## 2. RANCANGAN ANALISIS DAN ALGORITMA

*Feature Engineering and Selection*



**FEATURE AWAL**

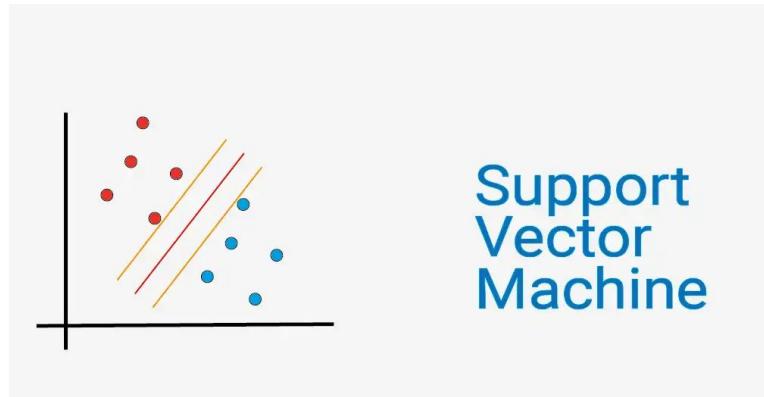
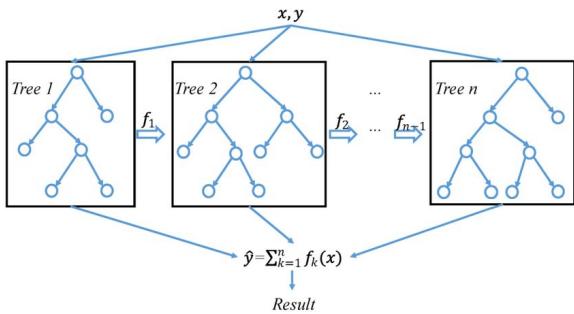
**FEATURE  
ENGINEERING**

**FEATURE  
SELECTION**

## 2. RANCANGAN ANALISIS DAN ALGORITMA

*Modelling*

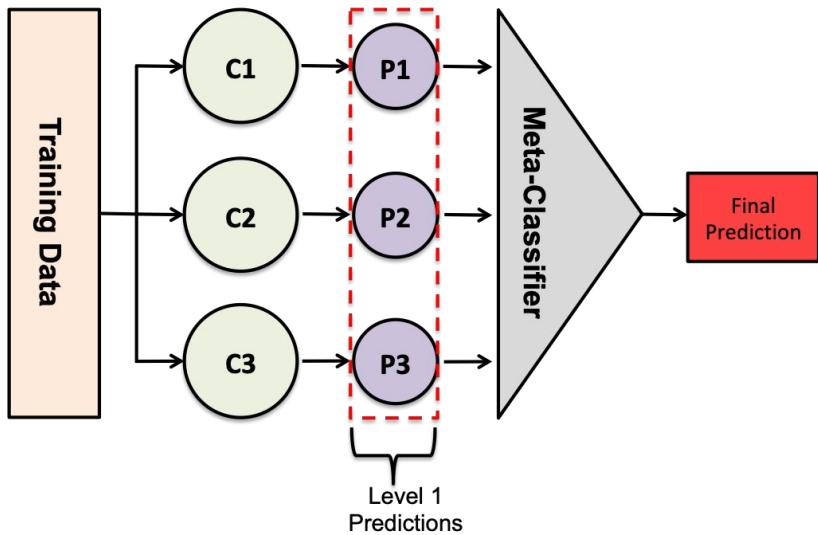
*dmlc*  
**XGBoost**



Support  
Vector  
Machine

## 2. RANCANGAN ANALISIS DAN ALGORITMA

Modelling : Stacking Classifier



- C1 : XGBoost Default
- C2 : XGBoost with scale\_post\_weight
- C3 : XGBoost with Bayesian Optimization
- Meta-Classifier : SVM

\* C1, C2, and C3 are considered level 1 classifiers.

## 2. RANCANGAN ANALISIS DAN ALGORITMA

*Modelling : Stacking Classifier*



```
xgb = XGBClassifier(tree_method='gpu_hist')
xgb2 = XGBClassifier(tree_method='gpu_hist',scale_pos_weight=1.85481664546)
xgb_tuned = XGBClassifier(colsample_bytree=0.49309299633453885,learning_rate=0.09082965940463038,
                           max_depth=10,min_child_weight=1,n_estimators=694,
                           reg_alpha=0.1,reg_lambda=1,
                           scale_pos_weight=4.6847053521598830,
                           subsample=1,
                           tree_method='gpu_hist')
```

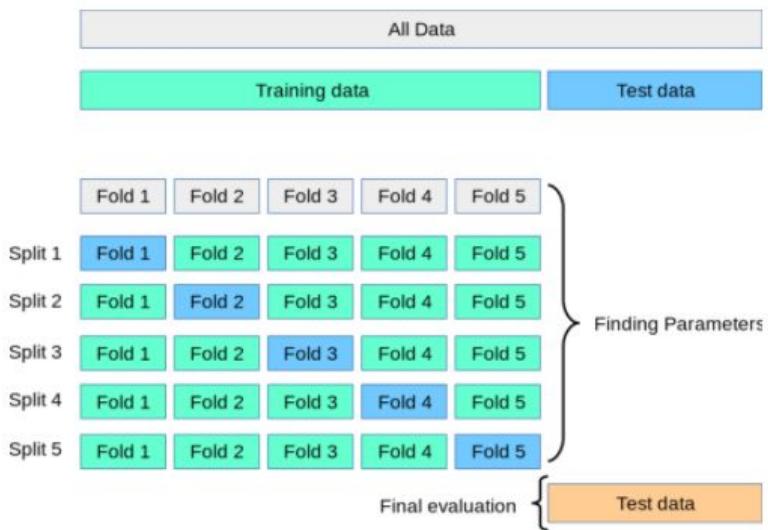


```
from mlxtend.classifier import StackingClassifier
from sklearn.svm import SVC

clf2 = StackingClassifier(classifiers =[xgb, xgb2, xgb_tuned], meta_classifier = SVC(kernel='linear',C=1, degree=3) ,
                           use_probas = True)
clf2.fit(X_train, y_train)
```

## 2. RANCANGAN ANALISIS DAN ALGORITMA

### Evaluation



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$



### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

Uji ANOVA untuk Kolom Numerik

Variabel	F Score	P Value
bobot	131801.917	0.000
tanggal	8390.253	0.000
lama_rawat	7507.107	0.000
hari	4609.151	0.000
quarter	847.399	2.794e-186
bulan	812.241	1.225e-178
pekanpadatahun	530.369	2.388e-177
haripadatahun	444.549	1.120e-98
tahun	13.584	2.280e-04



### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

Uji Chi-Square untuk Kolom Kategorik

Variabel	Chi Score	P Value
kode_diag	5.541e+09	0.000
kota_fkr	8.147e+07	0.000
kode_fdiag_icd	5.686e+07	0.000
poli_fktp	2.200e+07	0.000
jenis_fkr	2.025e+06	0.000
kepemilikan_fkr	1.793e+06	0.000
tipe_fkr	9.507e+05	0.000
kota_fktp	8.639e+05	0.000
kepemilikan_fktp	8.428e+05	0.000
provinsi_fkr	7.889e+05	0.000

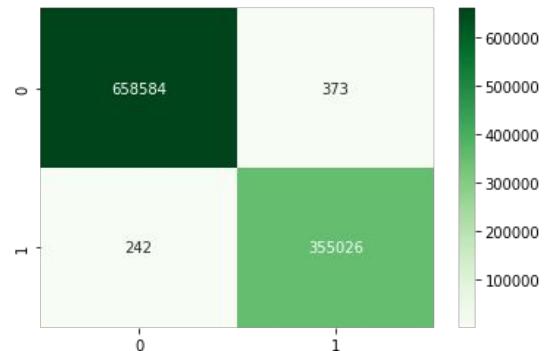
Variabel	Chi Score	P Value
jenis_kunjungan_fktp	6.489e+05	0.000
jenis_fktp	1.915e+05	0.000
segmen_peserta_fktp	1.477e+05	0.000
tipe_fktp	1.337e+05	0.000
tidak_dirujuk	2.112e+04	0.000
provinsi_fktp	8.440e+03	0.000
tingkatpel_fktp	7.424e+02	0.000
poli_fkr	1.952e+02	2.262e-44
is_weekend	1.212e+01	4.980e-04

### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

*Train-Test Split (75:25)*

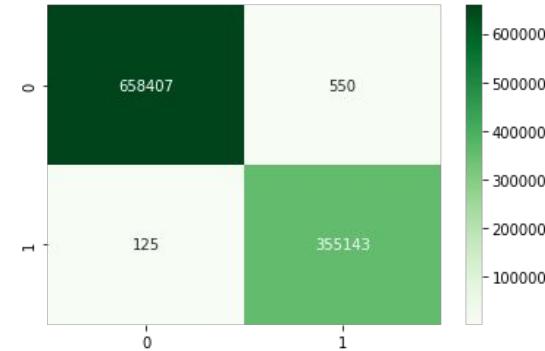
#### XGBOOST DEFAULT

F1 Score = 0.99913



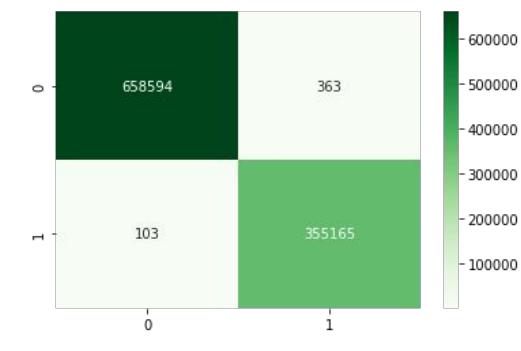
#### XGBOOST SCALE\_POS\_WEIGHT

F1 Score = 0.99905



#### XGBOOST BAYESIAN OPTIMIZATION

F1 Score = 0.99934



### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

*Train-Test Split (75:25)*

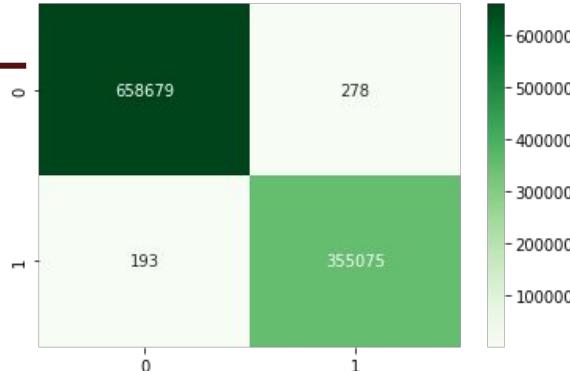
**XGBOOST DEFAULT**

**XGBOOST  
SCALE\_POS\_WEIGHT**

**XGBOOST BAYESIAN  
OPTIMIZATION**

**SVM**

F1 Score = 0.99933





### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

*Result*

XGBOOST DEFAULT

XGBOOST  
SCALE\_POS\_WEIGHT

XGBOOST BAYESIAN  
OPTIMIZATION

**SVM**

*Cross-Validation*

Avg F1 Score = 0.99932

*Submission (5th)*

F1 Score = 0.99939



### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

*F1 Score Performance*

<b>Model</b>	<b>F1 Score</b>
XGBoost	0.99913
XGBoost Scale_Pos_Weight	0.99905
XGBoost Tuned	0.99934
Stacking XGBoost + SVM	0.99933



### 3. HASIL ANALISIS, INSIGHT, DAN REKOMENDASI

*Running Time (GPU)*

Model	Train	Test
XGBoost	9.4 s	3.95 s
XGBoost Scale_Pos_Weight	5.78 s	4.56 s
XGBoost Tuned	28.5 s	20.8 s
Stacking XGBoost + SVM	110 s	25.5 s

## 4. Kesimpulan dan Saran

### Problem 1

01

Model klasifikasi untuk memprediksi status pulang pasien FKTP BPJS dengan skor F1 tertinggi dapat dibuat menggunakan algoritma *machine learning Stacking Classifier* (XGBoost + SVM). Performa model ini untuk rata-rata F1 Score pada validasi silang adalah 0.99932 dan pada submisi adalah 0.99939

02

Model ini dapat diterapkan oleh BPJS untuk memprediksi dari kunjungan pasien apakah status pulangnya sembuh atau tidak. Dengan demikian BPJS dapat melakukan *tracking* kepada pasien tersebut dan dapat melakukan pengefektifan anggaran dengan melihat status pulang pasien



# PROBLEM 2

**KLASTERISASI KLAIM DAN SENTIMEN  
BPJS KESEHATAN DI INDONESIA**

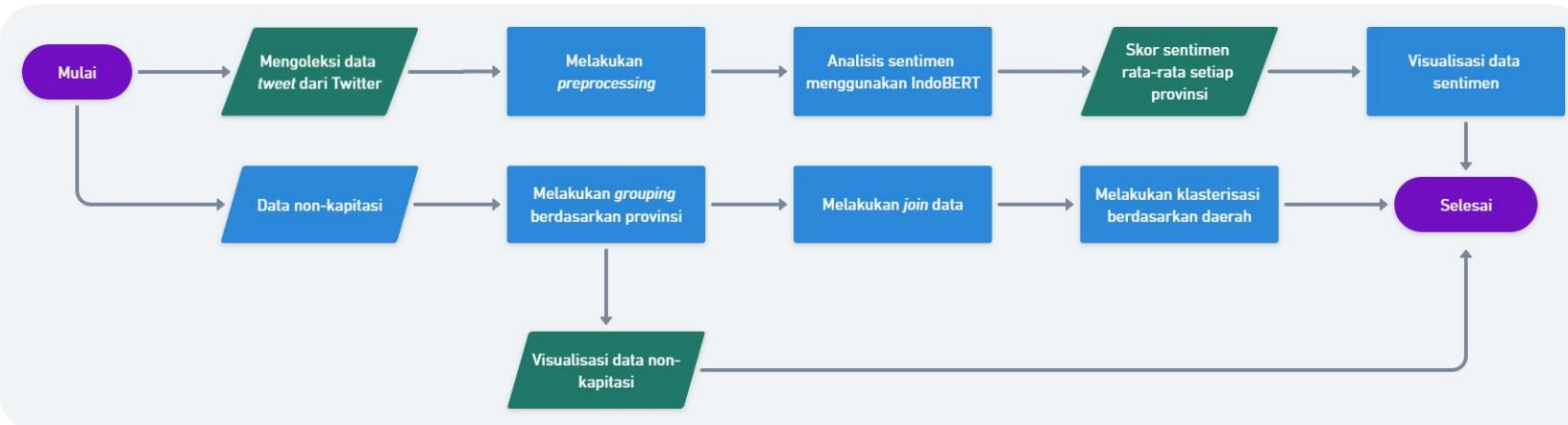
# Tujuan dari Kegiatan

Melihat sebaran karakteristik data non-kapitasi BPJS di setiap provinsi di Indonesia

Melihat sebaran sentimen pengguna Twitter di setiap provinsi di Indonesia

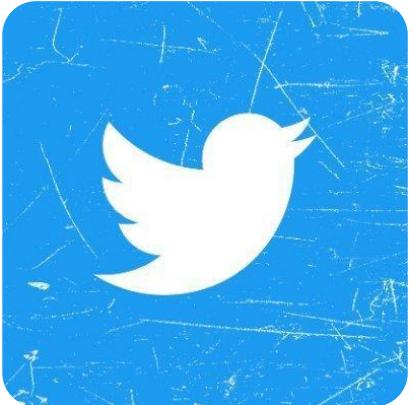
Menentukan klaster-klaster daerah tentang pelayanan BPJS di Indonesia

# METODOLOGI



# PROSES SCRAPING

## SUMBER DATA:



## MENGAPA TWITTER?

*Data tweet cukup merepresentasikan opini dari masyarakat*

*Data tweet dapat diakses secara publik (dengan bantuan scraping)*

# PROSES SCRAPING

**LIBRARY  
SCRAPING**

**SNSCRAPE**

Tidak memerlukan API dari Twitter secara langsung

Tidak memiliki batasan kuota dalam melakukan *scraping*



# PROSES SCRAPING

## QUERY

```
query = 'bpjs near:' + city + "  
within:20km since:2021-01-01  
until:2022-10-10 lang:id'
```

- “**bpjs**” adalah kata kunci dari pencarian tweet
- Variabel ‘**city**’ diisi dengan nama kota tiap provinsi
- Parameter ‘**within**’ merupakan radius dari pencarian
- Parameter ‘**since**’ dan ‘**until**’ digunakan untuk mengatur rentang waktu pencarian

# PREPROCESSING

Bertujuan menghapus:

**Tanda baca**

**Angka**

**Stopwords**

Serta melakukan:

**Stemming**

# SENTIMENT ANALYSIS

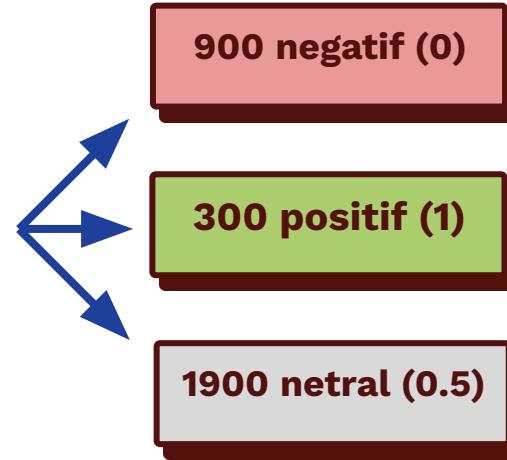
## MODEL

**INDOBERT**



**3141 tweet**

**IndoNLU**



# GROUPING PROVINSI DATA non-KAPITASI

Dilakukan grouping per provinsi dari data-data baris tiap kunjungan di berbagai daerah. Variabel yang dibentuk:

- **Jumlah peserta**
- **Rata-rata Lama Rawat**
- **Tipe faskes**
- **Jumlah KK**
- **Jumlah Kota**
- **Tingkat pelayanan**
- **Rata-rata bobot**
- **Kepemilikan faskes**
- **Segmentasi peserta**
- **Jumlah kunjungan**
- **Jenis faskes**
- **Diagnosa penyakit**
- **Nama tindakan**
- **Biaya klaim**

# EKSPLORASI KARAKTERISTIK DATA non-KAPITASI

## DATA JUMLAH PESERTA (SAMPEL)



>5.000

2.000 - 5.000

<2.000

# EKSPLORASI KARAKTERISTIK DATA non-KAPITASI

## DATA JUMLAH KUNJUNGAN BERDASARKAN WILAYAH (SAMPEL)



- >20.000
- 7.500 -  
20.000
- <7.500

# EKSPLORASI KARAKTERISTIK DATA non-KAPITASI

## DATA RATA-RATA LAMA RAWAT PER WILAYAH (HARI)



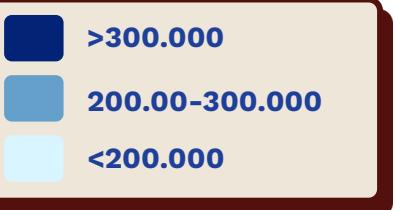
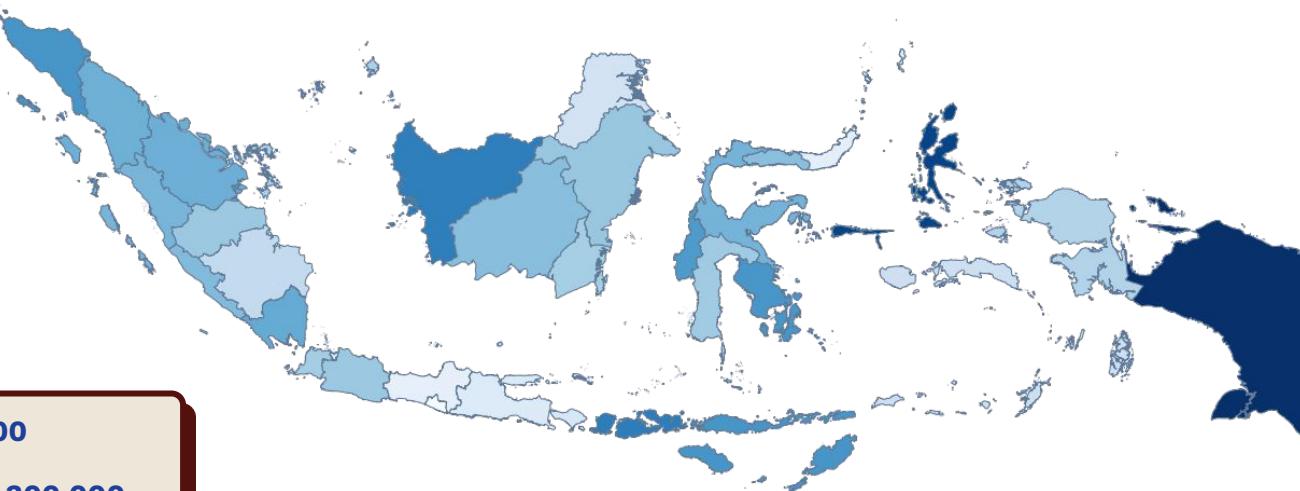
>0.35

0.15 - 0.35

<0.15

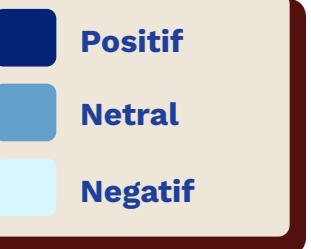
# EKSPLORASI KARAKTERISTIK DATA non-KAPITASI

## RATA-RATA BIAYA CLAIM (RUPIAH)



# EKSPLORASI KARAKTERISTIK DATA non-KAPITASI

RATA-RATA sentimen BERDASARKAN WILAYAH



# EKSPLORASI HASIL KLASTERISASI DAERAH

## KLASTERISASI DENGAN ALGORITMA K-means CLUSTERING



Klaster 1

Klaster 2

Klaster 3

Terbentuk 3 klaster dari hasil clustering:

### Klaster 1:

- Pulau Sumatera (kecuali Aceh, Sumatera Utara dan Lampung)
- DIY
- Pulau Kalimantan (kecuali Kalimantan Barat)
- Pulau Sulawesi (kecuali Sulawesi Selatan)
- Maluku Utara
- Maluku
- serta seluruh Provinsi yang ada di Pulau Papua.

# EKSPLORASI HASIL KLASTERISASI DAERAH

## KLASTERISASI DENGAN ALGORITMA K-means CLUSTERING



Klaster 1

Klaster 2

Klaster 3

Terbentuk 3 klaster dari hasil clustering:

### Klaster 2:

- Provinsi Jawa Barat
- Jawa Timur
- Jawa Tengah

# EKSPLORASI HASIL KLASTERISASI DAERAH

## KLASTERISASI DENGAN ALGORITMA K-means CLUSTERING



Klaster 1

Klaster 2

Klaster 3

Terbentuk 3 klaster dari hasil clustering:

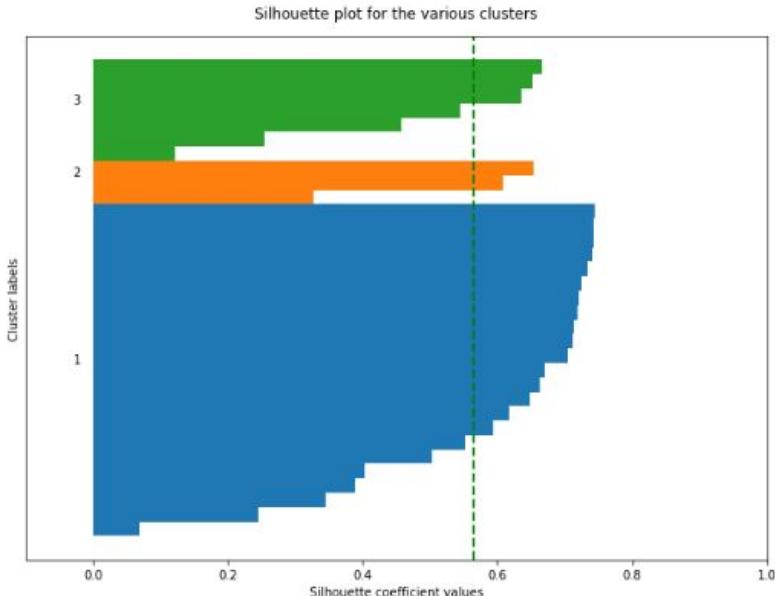
### Klaster 3:

- Provinsi Aceh
- Sumatera Utara
- Lampung
- Banten
- Kalimantan Barat
- Nusa Tenggara
- Sulawesi Selatan



# ANALISIS TAMBAHAN

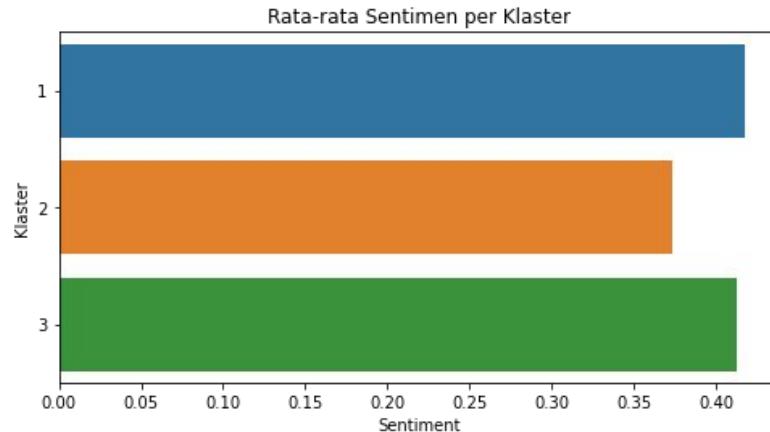
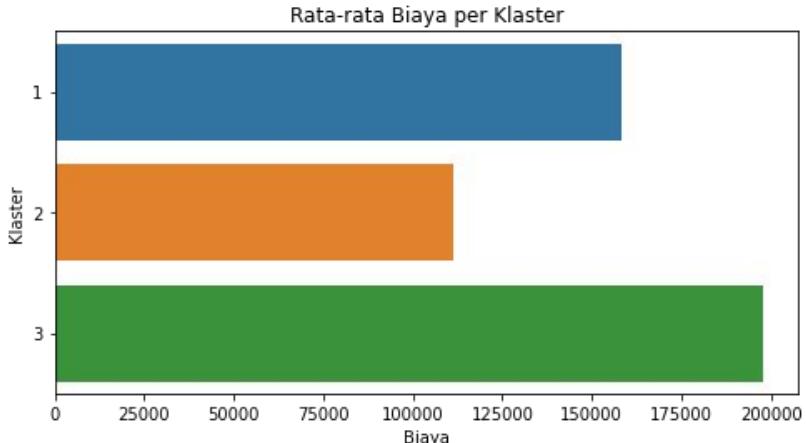
## PERFORMA KLASTERISASI (SILHOUTTE COEFFICIENT)



**SC = 0.58**

# ANALISIS TAMBAHAN

## KLASTER TERHADAP BIAYA DAN SENTIMEN



# KESIMPULAN

## Problem 2

01

Jumlah peserta serta pasien yang menggunakan BPJS masih terpusat di Pulau Jawa saja. Ini menjadi sinyal bagi pihak BPJS untuk segera melakukan penyuluhan pada daerah-daerah di luar Pulau Jawa, agar seluruh masyarakat Indonesia dapat menikmati dan terlindungi oleh fasilitas asuransi BPJS.



# KESIMPULAN

## Problem 2

02

Rata-rata **lama rawat** pasien BPJS paling besar ada di Provinsi Papua, Maluku, dan Nusa Tenggara.

03

Provinsi Papua, Maluku, dan Kalimantan Barat memiliki rata-rata **biaya klaim** terbesar. Banyak pengobatan yang memerlukan **biaya mahal**.

# KESIMPULAN

## Problem 2

04

**Hasil analisis sentimen:** Pulau Jawa dan Bali (dengan jumlah peserta serta kunjungan terbesar), Sumatera Barat, Sumatera Utara, Kalimantan Tengah, sebagian besar daerah Sulawesi serta Maluku Utara memiliki **sentimen yang mengarah ke negatif**.

05

Ketiga klaster yang terbentuk dapat digunakan untuk melakukan segmentasi daerah dalam tindakan yang dilakukan oleh BPJS. **Klaster 1 dan 3 berfokus pada biaya** dan **klaster 2 berfokus pada layanan**.

# TERIMA KASIH

# SD20220000074