INDUSTRIAL PRODUCTIVITY PREDICTION MODEL OF INDONESIA MEDIUM AND LARGE INDUSTRIES

Final Project Proposal Group 5 Data Science Bootcamp Shift Academy Batch 13

MEET THE TEAM

GROUP 5

01 Adi Permadi Jaya

02 Emir Khairy

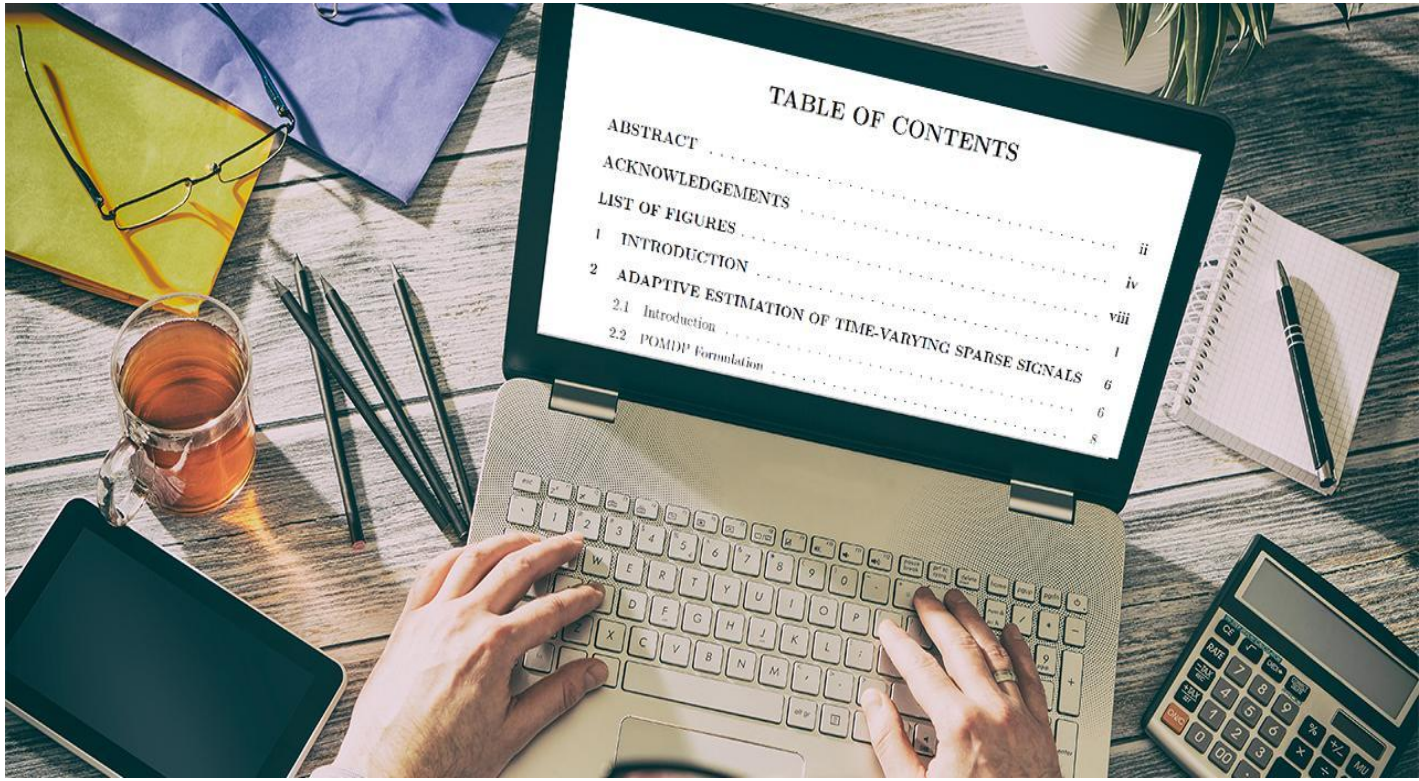
03 Muhammad Fikri A. T.

04 Rizalul Kalam

Coach: Faris Rizky Andika

TABLE OF CONTENTS

Table of Contents



BACKGROUND

PROJECT: INDUSTRIAL PRODUCTIVITY PREDICTION MODEL OF INDONESIA MEDIUM AND LARGE INDUSTRIES

Kontribusi sektor industri terhadap perekonomian nasional mencapai 19% (BPS, 2021).

Maka dari itu, peran industri besar dan menengah cukup penting karena merupakan sektor formal yang menyerap tenaga kerja dalam menggerakkan perekonomian nasional.

Oleh sebab itu, pemerintah perlu memperhatikan kinerja industri besar dan menengah untuk mengambil kebijakan-kebijakan yang tepat dan relevan.

GOALS

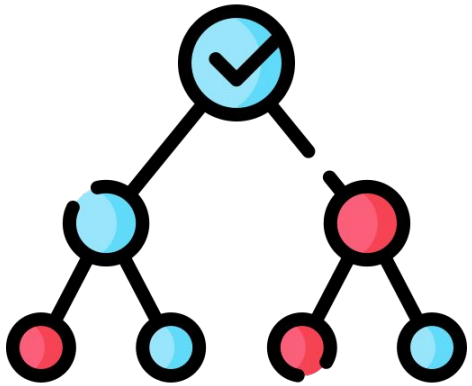
01 Business Goal

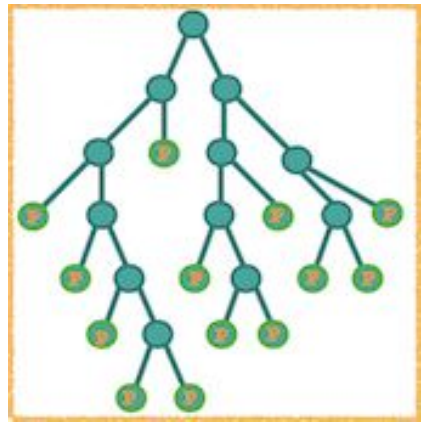
Untuk **menentukan perusahaan yang memiliki produktivitas tinggi** berdasarkan karakteristik individu perusahaan.

Berdasarkan informasi tersebut, pemerintah dapat menentukan kebijakan apa yang dapat diterapkan pada perusahaan yang memiliki produktivitas baik dan perusahaan yang tidak memiliki produktivitas yang baik.
Contoh: Insentif fiskal, Insentif non-fiskal.

02 Data Mining Goal

Membangun **model classifier** yang memiliki kemampuan untuk memprediksi apakah suatu perusahaan memiliki produktivitas yang tinggi atau tidak.





ABOUT THE DATA

01

Sumber Dataset: BPS Published Data

https://drive.google.com/drive/folders/1a6uuL_ZUEAdJIdXEAfqU-vik0G3cgyt7?usp=sharing

02

Jumlah Data

Jumlah sampel data sebanyak 27998 perusahaan (data training dan data test akan ditentukan kemudian)

03

Jumlah Variabel

1 target
129 variabel (hanya variabel yang mempengaruhi produktifitas yang akan digunakan)
Tipe Data: bool, integer, float

ABOUT THE DATA (CONT'D)

HIGH PROD	01	Company productivity, 1 : high productivity, 0 : low productivity
DSTATS	02	Company status, 1 : Domestic, 2 : Foreign, 3 : Non-facility
LOCATI	03	Company location, 1 : inside industrial estate, 2: outside industrial estate
DDOM	04	Percentage of capital owned by domestic (Govt, Local Govt, National Private Firm)
DASING	05	Percentage of capital owned by foreign (%)
DPROVI	06	Province
PRPRCA	07	Percentage of actual production to production capacity

Target yang akan diprediksi

Industri yang memperoleh fasilitas dan tidak

Teori lokasi

Teori kepemilikan dimana perusahaan asing lebih produktif (Crespo & Fontoura, 2007)

Teori lokasi

Skala Ekonomi

ABOUT THE DATA (CONT'D)

DISIC5	08	Industrial Sub Sector (F&B, Textile, Automotive, Chemical, etc)
V1115	09	Capital
EKSPOR	10	Export, 1 : Export, 2: No Export
LPMLTL	11	Male production labor
LPWLTL	12	Female production labor
LNWLTL	13	Female Non production labor
LNMLTL	14	Male Non production labor

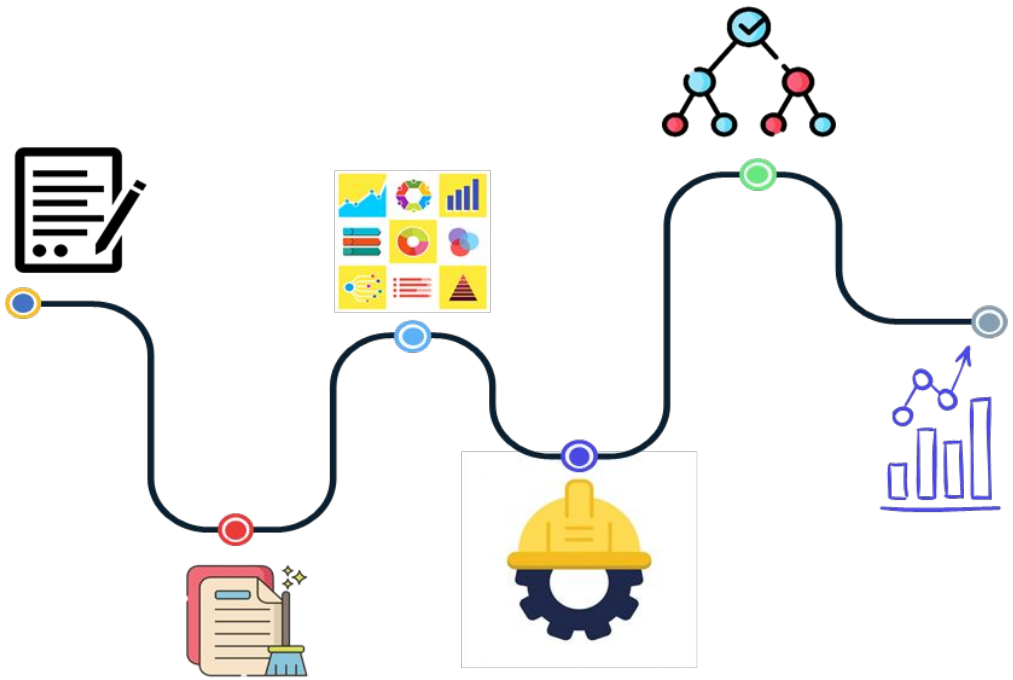
Melihat jenis industri yang lebih produktif

Teori pertumbuhan sollow (Capital, Labor)

Dugaan kuat bahwa perusahaan yang ekspor lebih produktif

Modifikasi: Perbandingan jumlah pekerja di produksi dibandingkan total jumlah pekerja

PROJECT FLOW STRATEGY



Topic & Dataset

- Determining which type of ML model to make (regression, classification, or clustering)
- Selecting an open source dataset
- Creating a proposal presentation

Data Cleaning

- Reformatting data
- Handling null and missing values
- Dropping data duplicates
- Handling outlier

EDA

- Creating data visualizations to gain insights

Featuring Engineering

- Scaling features
- Encoding categorical features
- Reducing data dimensions (PCA)
- Handling imbalance classes
- Selecting 'important' feature

Modelling

- Testing out various classification algorithms with training data
- Evaluating the performance metrics (recall, precision, f1 score, ROC/AUC

Model Evaluation

- Predicting test data using model which has the best performance metrics
- If the result is not good enough, go back to feature engineering / modelling phase

Understanding Data

Look out our data

```
In [2]: df1 = pd.read_csv('DATA_PROJECT.csv')
        df1.head()
```

Out[2]:

	DPROVI07	DKABUP07	DSTATS07	LOCATI07	JAN07	FEB07	MAR07	APR07	MEI07	JUN07	JUL07	AGS07	SEP07	OKTC
0	11	15.0	2	NaN	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
1	11	15.0	3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
2	11	14.0	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
3	11	14.0	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
4	11	14.0	1	NaN	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1

Data set tersebut terdiri dari 130 kolom dan 27.998 baris. Namun, hanya 14 kolom yang difilter untuk proses analisa.

Filtering columns

```
[6]  df1['PPL']=(df1['LPMLTL07']+df1['LPWLTL07'])/df1['LTLNOU07']
```

Menambahkan kolom baru 'Percentage of Production Labor' ['PPL'] yang merupakan penambahan dari kolom 'LPMLTL07' (Male Production Labor) dan kolom 'LPWLTL07' (Female Production Labor)

Filtering columns

```
In [53]: df['DDOM07'] = df['DPUSAT07']+df['DPEMDA07']+df['DDMSTK07']
```

Menambahkan kolom baru 'Percentage of Company Ownership by Domestic Parties' ['DDOM07'] yang merupakan penambahan dari kolom 'DPUSAT07' (Percentage of Company Ownership by the Central Government). 'DPEMDA07' (Percentage of Company Ownership by Regional Governments) dan 'DDMSTK07' (Percentage Ownership of the Company by local Parties)

Filtering columns

In [52]:
```
# filter kolom yang digunakan untuk analisa
df = df1[['PSID','DPROVI07','DSTATS07', 'LOCATI07', 'DISIC507','DASING07','EKSPOR07','PRPRCA07',
          'V1115','PPL','HIGH PROD']]
df.head()
```

Out[52]:

```
# kolom yang digunakan untuk proses cleansing data adalah sebanyak 12 kolom
df = df.drop(columns=['DPUSAT07','DPEMDA07','DDMSTK07'])
df.head()
```

	PSID	DPROVI07	DSTATS07	LOCATI07	DISIC507	DASING07	EKSPOR07	PRPRCA07	V1115	PPL	HIGH PROD	DDOM07
0	1761	11	2	NaN	15141	90.0	NaN	50.0	0.0	0.616279	0	10.0
1	1762	11	3	1.0	15141	90.0	2.0	0.0	0.0	0.099910	0	10.0
2	1763	11	1	NaN	15141	0.0	NaN	0.0	0.0	0.973684	0	100.0
3	1765	11	1	NaN	15141	0.0	NaN	0.0	0.0	0.750000	0	100.0
4	1766	11	1	NaN	15141	0.0	NaN	0.0	0.0	0.564516	1	100.0

Selanjutnya, kolom final adalah sebanyak 12 kolom yang akan diolah untuk proses cleansing data dengan kolom ‘HIGH PROD’ sebagai kolom target.

Filtering columns

Kolom akhir yang digunakan:

No.	Nama Kolom	Arti Kolom
1	PSID	ID Perusahaan
2	DPROVI07	Wilayah dalam Provinsi
3	DSTATS	Status Perusahaan (1.Domestik, 2.Asing, 3.Non-Facility)
4	LOCATI	1: Di dalam kawasan industri. 2: Di luar kawasan industri
5	DISIC507	Klasifikasi Jenis Industri
6	DASING07	Persentase Kepemilikan Modal oleh Asing
7	DDOM07	Persentase Kepemilikan Modal oleh Lokal
8	EKSPOR07	Apakah ada produksi yang diekspor. 1:Ya, 2:Tidak
9	PRPRCA07	Persentasi dari realisasi produksi terhadap kapasitas terpasang
10	V1115	Jumlah total modal usaha
11	PPL	Persentase jumlah tenaga kerja produksi dibandingkan dengan keseluruhan tenaga kerja produksi
12	HIGH PROD	Produktifitas perusahaan (1:HIGH, 0: LOW)

EDA | Duplicated & Outlier

```
In [7]: # check duplicat
        df.duplicated().any().sum()
```

```
Out[7]: 0
```

```
Out[56]: <AxesSubplot:xlabel='PRPRCA07'>
```

```
In [57]: df=df[df['PRPRCA07']<=100]
```

PRPRCA07 = Persentasi dari realisasi produksi terhadap kapasitas terpasang

EDA | Missing Value

```
In [67]: #menghapus modal di quantil 25
         df=df[df['V1115']>=170000]
         df['Modal Dalam Jutaan Rupiah'] = df['V1115']/1000000
```

Menghapus data di quantil 25 karena menurut kami itu data tersebut rancu. Lalu, menambahkan kolom baru 'Modal Dalam Jutaan Rupiah'.

EDA | Missing Value

```
In [63]: # check presentasi Missing value
         df.isnull().sum()/df.shape[0]*100
```

```
Out[63]: PSID          0.000000
         DPROVI07      0.000000
         DSTATS07      0.000000
         LOCATI07     33.316330
         DISIC507      0.000000
         DASING07      0.000000
         EKSPOR07     33.396488
         PRPRCA07     14.147781
         V1115        47.784741
         PPL           0.000000
         HIGH PROD     0.000000
         DDOM07        0.003644
         dtype: float64
```

```
In [66]: df = df.dropna(axis=0)
         df
```

Karena persentase missing value lebih dari 30%, maka kami memilih untuk menghapus seluruh data missing.

EDA | Data Type & Feature Transformation

```
[25] df['HIGH PROD'].replace('0',0,inplace=True)
     df['HIGH PROD'].replace('1',1,inplace=True)
```

HIGH PROD = Produktifitas perusahaan (0: Low, 1:High,)

```
# Jawa = 1, luar jawa = 0
df1['DPROVI07'][(df1['DPROVI07']<30)]=0
df1['DPROVI07'][(df1['DPROVI07']>40)]=0
df1['DPROVI07'][(df1['DPROVI07']>30)&(df1['DPROVI07']<40)]=1
```

DPROVI07 = Lokasi perusahaan dalam provinsi (0. Luar Jawa, 1. Jawa)

EDA | Data Type & Feature Transformation

Dilakukan encode untuk variabel DISIC507 (Jenis Industri) agar mudah untuk mengetahui pengaruh dari jenis industri terhadap produktivitas perusahaan.

```
In [80]: df1['DISIC507'] = df1['DISIC507'].astype('int')
```

```
In [81]: #mengubah kode jenis industri menjadi lebih simple (dikelompokan menjadi 8 jenis industri lihat link data)
         df1['DISIC507'][(df1['DISIC507']>0)&(df1['DISIC507']<290)]=6
         df1['DISIC507'][(df1['DISIC507']>289)&(df1['DISIC507']<15000)]=7
         df1['DISIC507'][(df1['DISIC507']>14999)&(df1['DISIC507']<17000)]=1
         df1['DISIC507'][(df1['DISIC507']>16999)&(df1['DISIC507']<20000)]=2
         df1['DISIC507'][(df1['DISIC507']>19999)&(df1['DISIC507']<22000)]=3
         df1['DISIC507'][(df1['DISIC507']>21999)&(df1['DISIC507']<23000)]=4
         df1['DISIC507'][(df1['DISIC507']>22999)&(df1['DISIC507']<24000)]=5
         df1['DISIC507'][(df1['DISIC507']>23999)&(df1['DISIC507']<25000)]=6
         df1['DISIC507'][(df1['DISIC507']>24999)&(df1['DISIC507']<26000)]=1
         df1['DISIC507'][(df1['DISIC507']>25999)&(df1['DISIC507']<34000)]=7
         df1['DISIC507'][(df1['DISIC507']>33999)&(df1['DISIC507']<36000)]=8
         df1['DISIC507'][(df1['DISIC507']>35999)&(df1['DISIC507']<38000)]=1
```

1 = industri agro
2 = industri tekstil
3 = industri pulp dan kertas
4 = industri percetakan
5 = industri pertambangan
6 = industri kimia
7 = industri logam dan mesin
8 = industri otomotif

EDA | Data Type & Feature Transformation

```
[39] df1['LOCATI07'][(df1['LOCATI07']==1)]=1
     df1['LOCATI07'][(df1['LOCATI07']==2)]=0
```

LOCATI07 = Lokasi perusahaan dalam dalam kawasan industri (0. Luar kawasan industri, 1. Dalam kawasan industri)

```
[40] df1['EKSPOR07'][(df1['EKSPOR07']==1)]=1
     df1['EKSPOR07'][(df1['EKSPOR07']==2)]=0
```

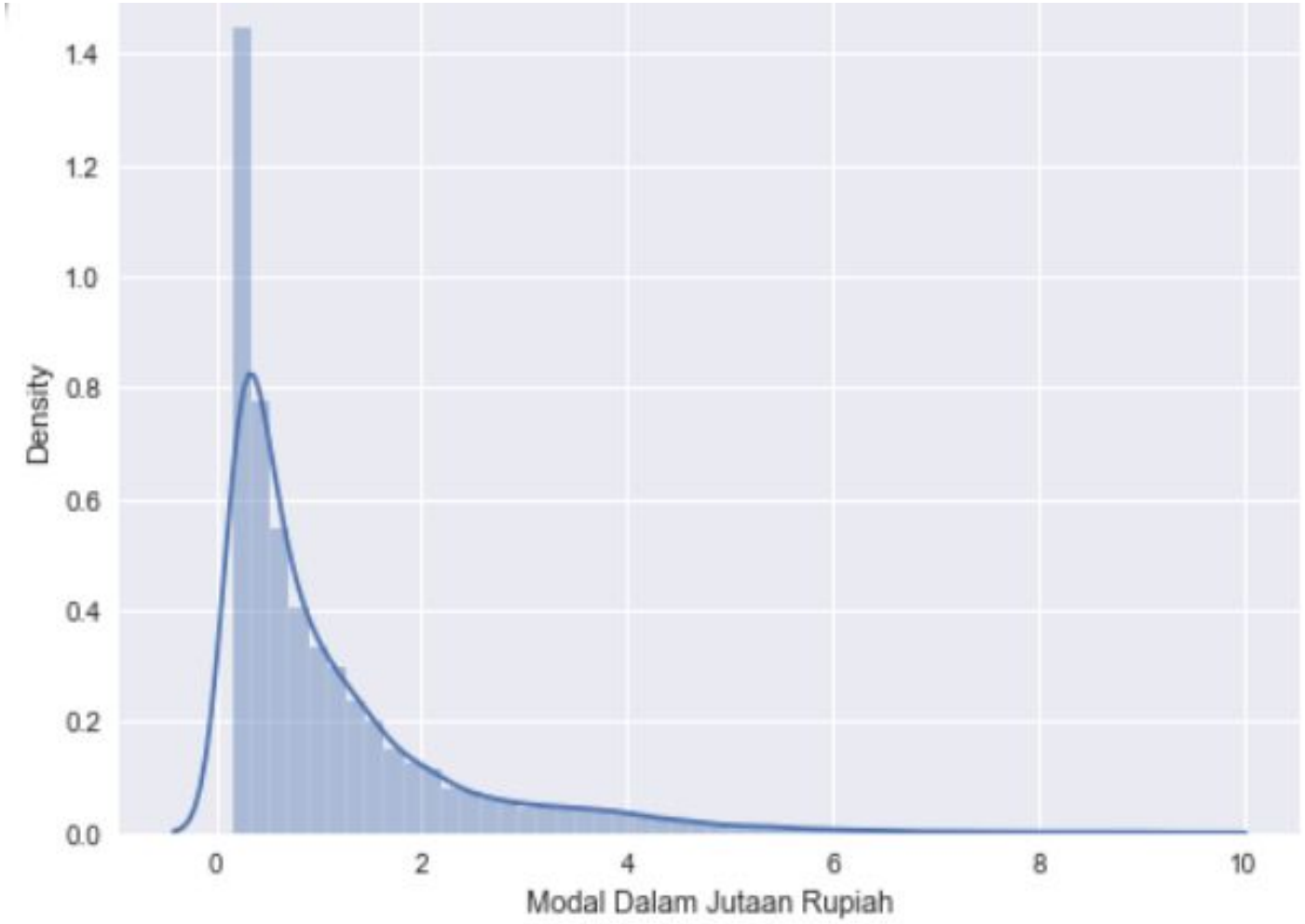
EKSPOR07 = Apakah ada produksi yang diekspor? (0. Tidak ada ekspor, 1. Ada sebagian atau Seluruh Produk yang diekspor)

Visualization

Klik tautan di bawah untuk melihat visualisasi di tableau.

https://public.tableau.com/views/FinalProject2_16473593909990/Sheet11?:language=en-US&:display_count=n&:origin=viz_share_link

Scaling

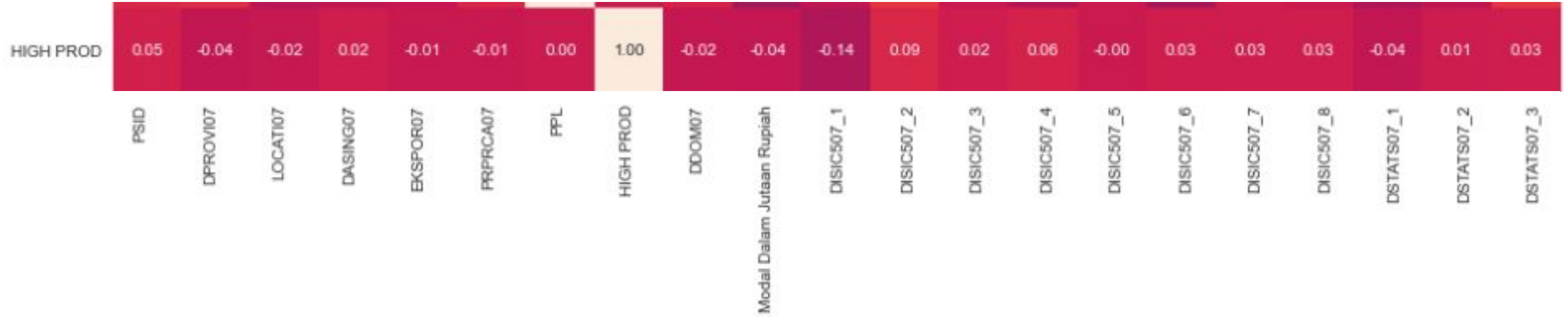


Variabel Modal (‘Modal Dalam Jutaan Rupiah’) distribusinya tidak normal (scew), sehingga digunakan scaling logaritma untuk mendapatkan hasil yang lebih baik.

df1['Modal Dalam Jutaan Rupiah']=np.log1p(df1['Modal Dalam Jutaan Rupiah'])

Correlation

```
In [44]:
plt.figure(figsize=(20, 20))
sns.heatmap(df1.corr(), annot=True, fmt='.2f')
```



Berdasarkan heatmap korelasi antara target dengan feature, dapat dilihat bahwa nilai korelasinya relatif kecil. Hal ini kemungkinan akan berpengaruh pada hasil pemodelan yang kurang memuaskan.

Modeling | Model preparation and Train Test Split

Kami membagi variabel X dan Y, lalu kami membagi lagi untuk data train dan data test.

In [301]:
```
x = df1.drop(['HIGH PROD'],axis = 1)
y = df1['HIGH PROD']
```

In [302]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size = 0.2, random_state = 42)
```

In [303]:
```
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(7889, 18)
(1973, 18)
(7889,)
(1973,)
```

- Menghapus PSID dan DASING07
- 20% dataset untuk data test

Modeling | SMOTE & UNDERSAMPLING

SMOTE

In [304]:
```
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
```

In [306]:
```
smote = SMOTE(random_state=1)
under_sampling = RandomUnderSampler(random_state=41)
```

In [307]:
```
y_train.value_counts()
```

Out[307]:
```
0    5302
1    2587
Name: HIGH PROD, dtype: int64
```

In [308]:
```
x_smote, y_smote = smote.fit_resample(x_train, y_train)
```

In [309]:
```
x_under, y_under = under_sampling.fit_resample(x_train, y_train)
```

In [310]:
```
y_smote.value_counts()
```

Out[310]:
```
0    5302
1    5302
Name: HIGH PROD, dtype: int64
```

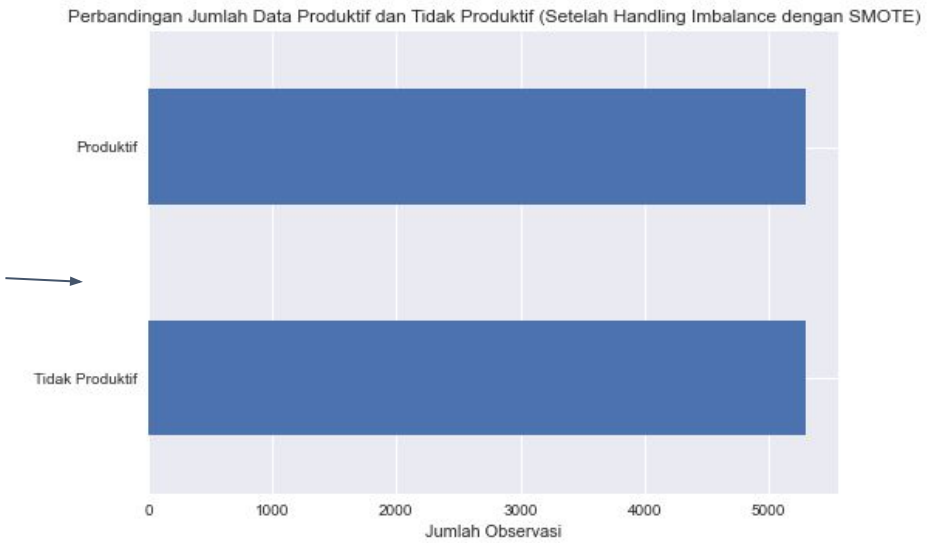
In [311]:
```
y_under.value_counts()
```

Out[311]:
```
0    2587
1    2587
Name: HIGH PROD, dtype: int64
```

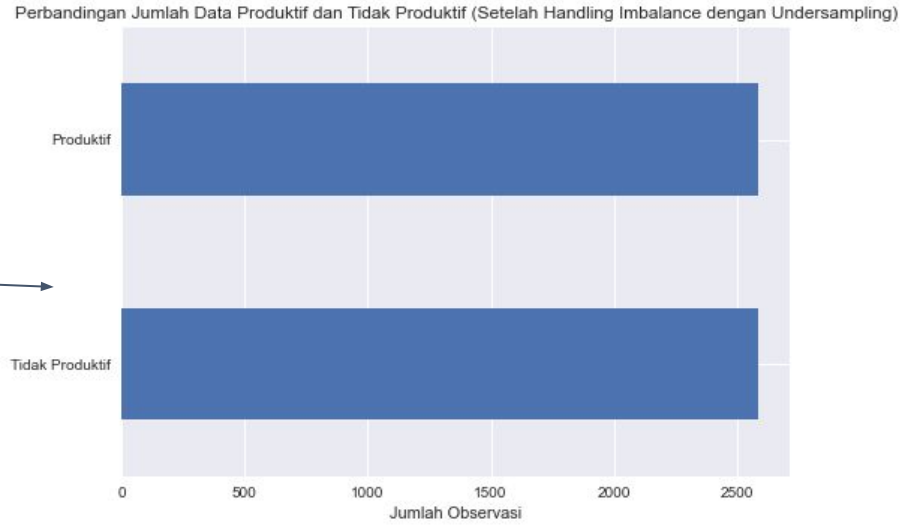
Modeling | SMOTE

Sebelum SMOTE dan Undersampling

Setelah SMOTE



Setelah Undersampling



Modeling | Modeling

Modeling

In [314]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
```

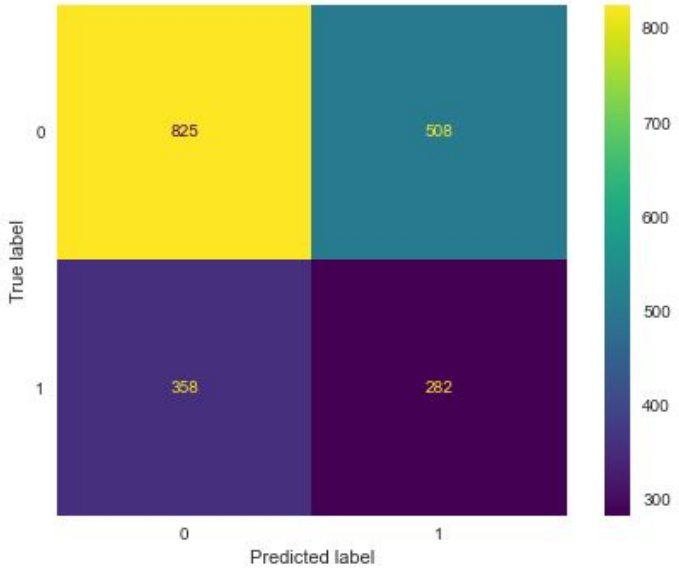
In [315]:
```
knn = KNeighborsClassifier(n_neighbors = 99)
rf = RandomForestClassifier(random_state = 123)
svm = SVC()
lr = LogisticRegression(random_state = 123)
dt = DecisionTreeClassifier(random_state = 123)
```

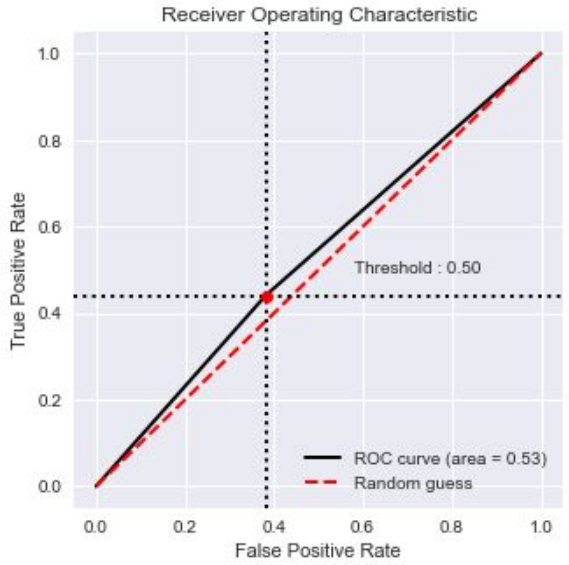
In [316]:
```
models = [rf, svm, lr, knn, dt]

for model in models:
    model.fit(x_train, y_train)
```

In [317]:
```
ypred_rf = rf.predict(x_test)
ypred_svm = svm.predict(x_test)
ypred_lr = lr.predict(x_test)
ypred_knn = knn.predict(x_test)
ypred_dt = dt.predict(x_test)
```

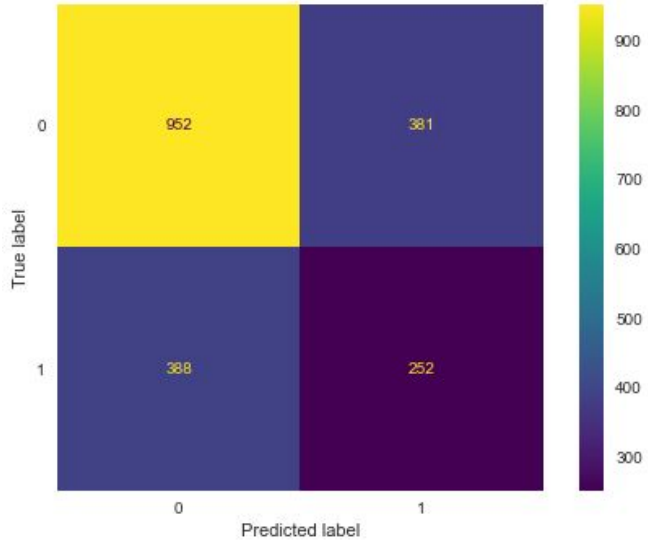
Modeling | Modeling

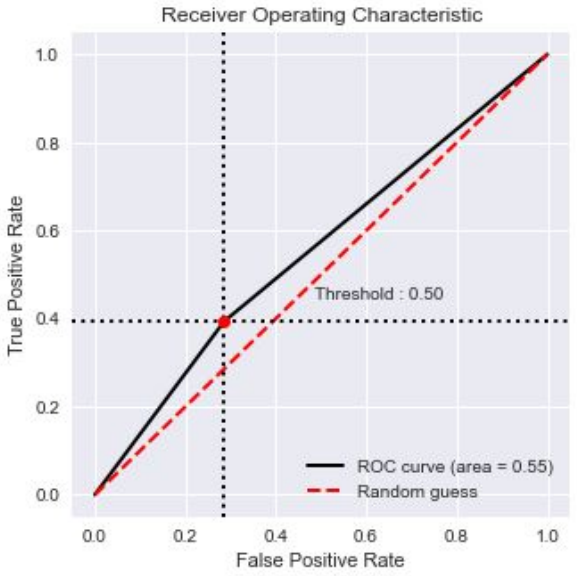




SMOTE Decision Tree

SMOTE Random Forest





Modeling | Evaluation

	LR Model Score (%)	DT Model Score (%)	RF Model Score (%)	SVM Model Score (%)	KNN Model Score (%)	LR Undersampling Model Score (%)	DT Undersampling Model Score (%)	RF Undersampling Model Score (%)	SVM Undersampling Model Score (%)	KNN Undersampling Model Score (%)
accuracy	67.663457	57.577293	65.129245	67.562088	67.359351	60.466295	56.107451	61.023822	63.507349	57.729346
recall	2.500000	34.531250	25.468750	0.000000	1.718750	45.937500	44.062500	39.375000	13.281250	52.968750
precision	53.333333	34.585290	43.582888	0.000000	42.307692	40.384615	35.696203	39.810427	34.000000	38.876147
roc_auc_score	50.724869	51.586705	54.819896	50.000000	50.296734	56.689680	52.976486	55.396427	50.451578	56.491877
f1_score	4.776119	34.558249	32.149901	0.000000	3.303303	42.982456	39.440559	39.591516	19.101124	44.841270

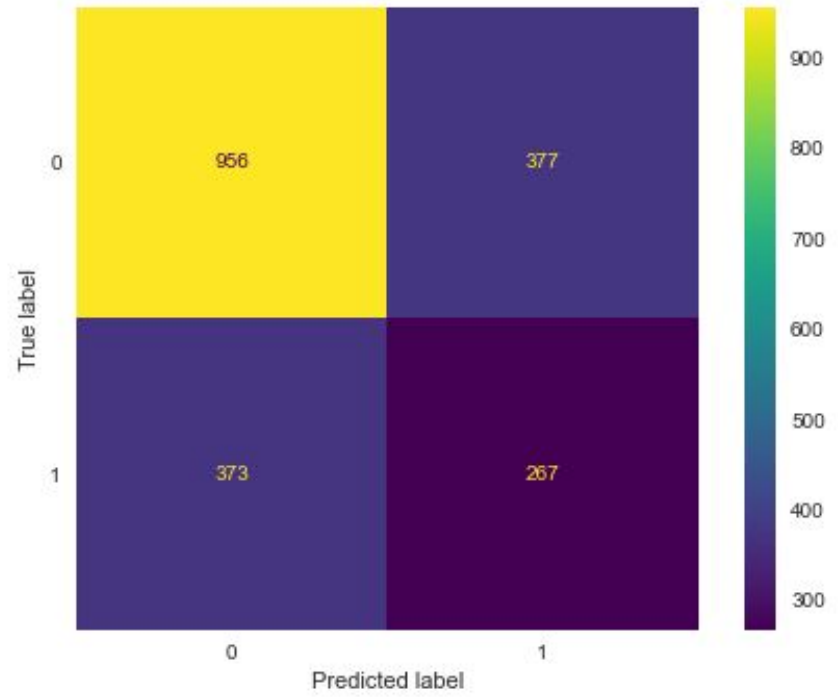
KNN Undersampling Model Score (%)	LR Smote Model Score (%)	DT Smote Model Score (%)	RF Smote Model Score (%)	SVM Smote Model Score (%)	KNN Smote Model Score (%)
57.729346	60.466295	56.107451	61.023822	63.507349	57.729346
52.968750	45.937500	44.062500	39.375000	13.281250	52.968750
38.876147	40.384615	35.696203	39.810427	34.000000	38.876147
56.491877	56.689680	52.976486	55.396427	50.451578	56.491877
44.841270	42.982456	39.440559	39.591516	19.101124	44.841270

Modeling | Tuning

	RF Smote Model Score (%)	RF Model Score After Tuning(%)
accuracy	61.023822	61.986822
recall	39.375000	41.718750
precision	39.810427	41.459627
roc_auc_score	55.396427	56.718340
f1_score	39.591516	41.588785

Score Result

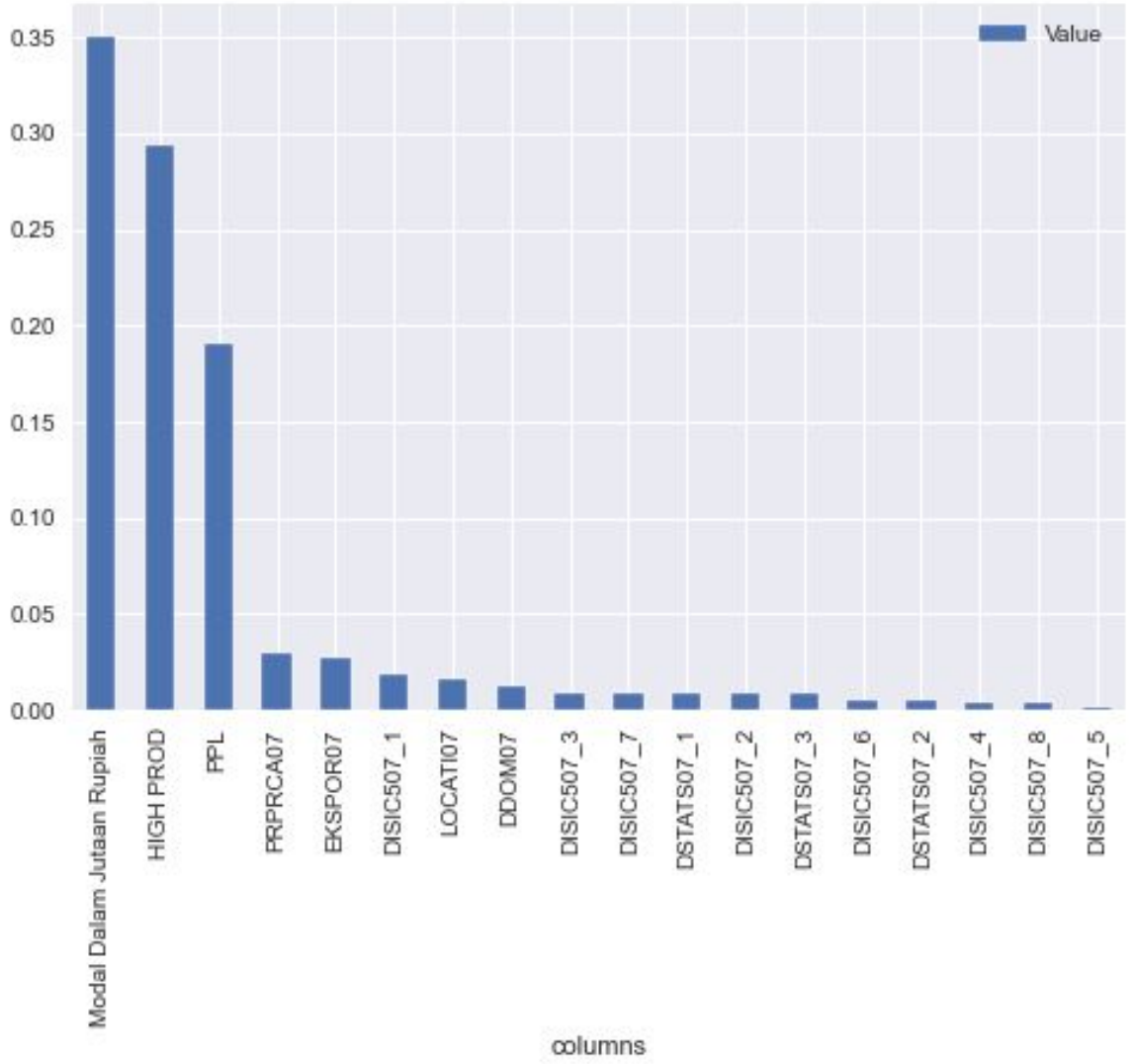
	precision	recall	f1-score	support
0	0.72	0.72	0.72	1333
1	0.41	0.42	0.42	640
accuracy			0.62	1973
macro avg	0.57	0.57	0.57	1973
weighted avg	0.62	0.62	0.62	1973



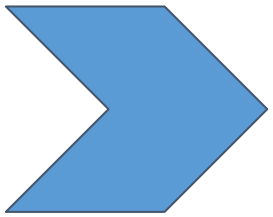
After tuning

Modeling | Feature Importance

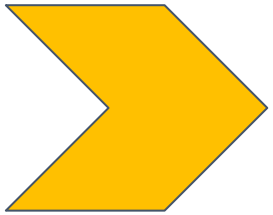
	columns	Importance	Value
0	LOCATI07	0.015740	0.015740
1	EKSPOR07	0.027530	0.027530
2	PRPRCA07	0.029426	0.029426
3	PPL	0.190406	0.190406
4	HIGH PROD	0.293566	0.293566
5	DDOM07	0.012788	0.012788
6	Modal Dalam Jutaan Rupiah	0.349665	0.349665
7	DISIC507_1	0.018907	0.018907
8	DISIC507_2	0.008347	0.008347
9	DISIC507_3	0.009140	0.009140
10	DISIC507_4	0.003919	0.003919
11	DISIC507_5	0.000913	0.000913
12	DISIC507_6	0.005278	0.005278
13	DISIC507_7	0.008848	0.008848
14	DISIC507_8	0.003443	0.003443
15	DSTATS07_1	0.008662	0.008662
16	DSTATS07_2	0.005138	0.005138
17	DSTATS07_3	0.008283	0.008283



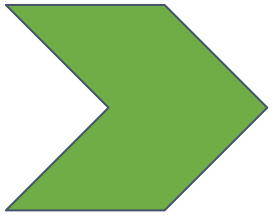
Future Plan



Menggunakan model lain seperti deep learning (jika memungkinkan) untuk mendapatkan hasil prediksi yang lebih baik



Mencari faktor lain yang berpengaruh terhadap produktivitas perusahaan



Melakukan merging data apabila terdapat data set lain yang memiliki ID perusahaan yang sama

Thank You

Shift academy