Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

LAPORAN

Tugas Besar MK Penambangan Data semester 2020-1

Task

Buat model untuk mengklasifikasi transaksi *fraud* atau tidak *fraud*

Terdapat dua file “train.csv” dan “test.csv” yang fitur-fiturnya *identical*.

File train.csv berisi data untuk melatih model klasifikasi, dan file test.csv digunakan untuk mencari prediksi *fraud*-nya

Kolom *fraud* berisi dua kemungkinan integer, (0,1)

Data Analisis

Analisis Fitur

Column name	Description	Value range
trustLevel	A customer’s individual trust level. 6: Highest trustworthiness	{1,2,3,4,5,6}
totalScanTimeInSeconds	Total time in seconds between the first and last product scanned	Positive whole number
grandTotal	Grand total of products scanned	Positive decimal number with maximum two decimal places
lineItemVoids	Number of voided scans	Positive whole number
scansWithoutRegistration	Number of attempts to activate the scanner without actually scanning anything	Positive whole number or 0
quantityModification	Number of modified quantities for one of the scanned products	Positive whole number or 0
scannedLineItemsPerSecond	Average number of scanned products per second	Positive decimal number
valuePerSecond	Average total value of scanned products per second	Positive decimal number
lineItemVoidsPerPosition	Average number of item voids per total number of all scanned and not cancelled products	Positive decimal number
fraud	Classification as fraud (1) or not fraud (0)	{0,1}

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

Isi data train

	trustLevel	totalScanTimeInSeconds	grandTotal	lineItemVoids	scansWithoutRegistration	quantityModifications	scannedLineItemsPerSecond	valuePerSecond	lineItemVoidsPerPosition	fraud
0	5	1054	54.70	7	0	3	0.027514	0.051898	0.241379	0
1	3	108	27.36	5	2	4	0.129630	0.253333	0.357143	0
2	3	1516	62.16	3	10	5	0.008575	0.041003	0.230769	0
3	6	1791	92.31	8	4	4	0.016192	0.051541	0.275862	0
4	5	430	81.53	3	7	2	0.062791	0.189605	0.111111	0

Isi data test

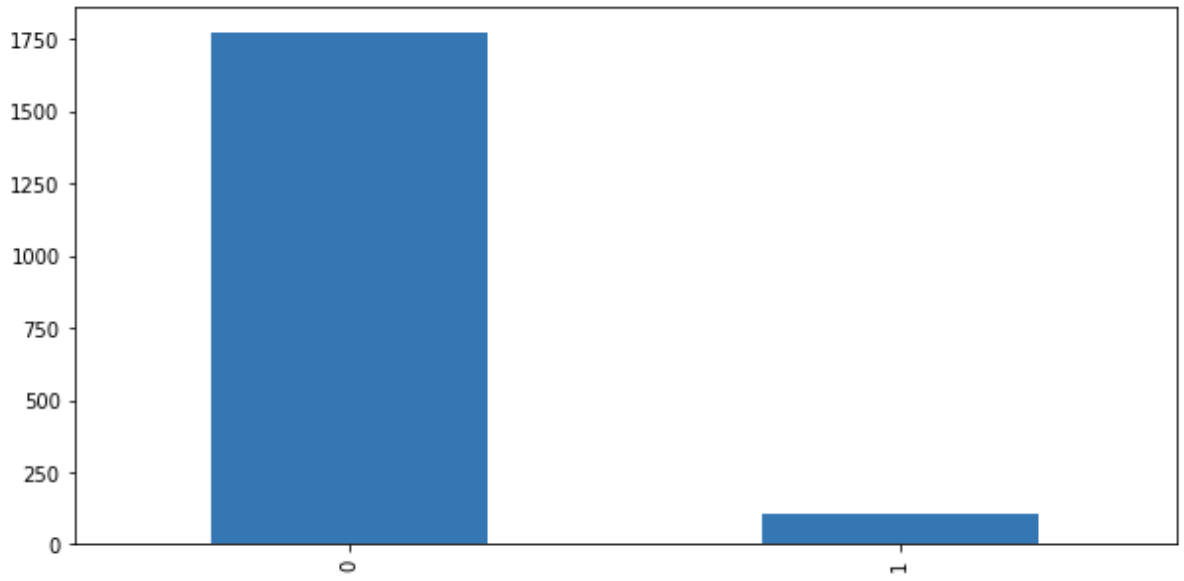
	trustLevel	totalScanTimeInSeconds	grandTotal	lineItemVoids	scansWithoutRegistration	quantityModifications	scannedLineItemsPerSecond	valuePerSecond	lineItemVoidsPerPosition
0	4	467	88.48	4	8	4	0.014989	0.189465	0.571429
1	3	1004	58.99	7	6	1	0.026892	0.058755	0.259259
2	1	162	14.00	4	5	4	0.006173	0.086420	4.000000
3	5	532	84.79	9	3	4	0.026316	0.159380	0.642857
4	5	890	42.16	4	0	0	0.021348	0.047371	0.210526

Tidak ada value yang kosong pada kedua dataset

```
print(df.isnull().values.any())
print(data_test.isnull().values.any())
```

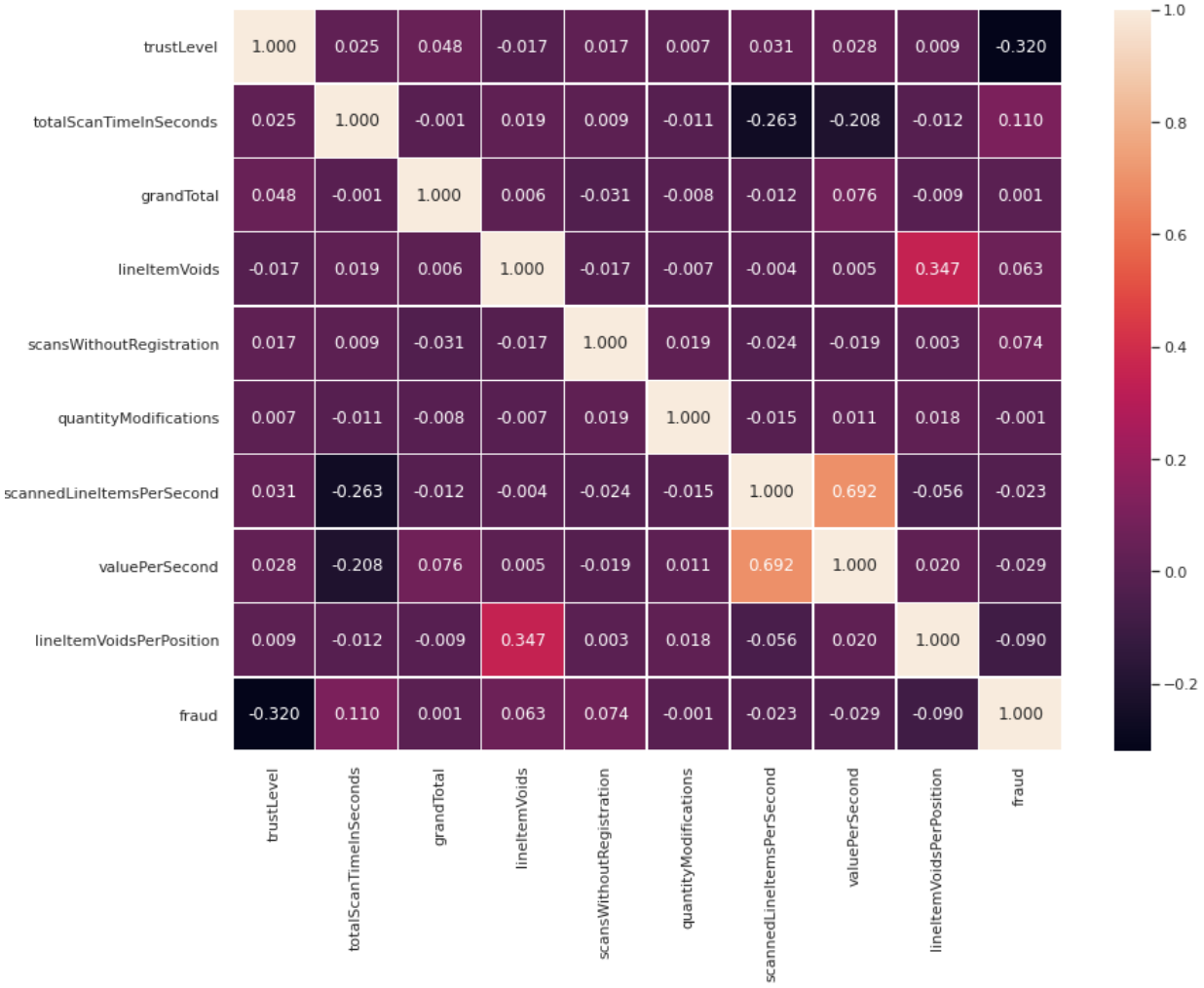
```
False
False
```

Distribusi data terhadap *fraud*



Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

Korelasi setiap fitur pada data train



Korelasi antar fitur terhadap fraud

```
fraud                        1.000000
trustLevel                   0.319765
totalScanTimeInSeconds       0.110414
lineItemVoidsPerPosition     0.090116
scansWithoutRegistration     0.074123
lineItemVoids                0.063496
valuePerSecond               0.028873
scannedLineItemsPerSecond    0.023085
grandTotal                   0.001421
quantityModifications        0.000864
```

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

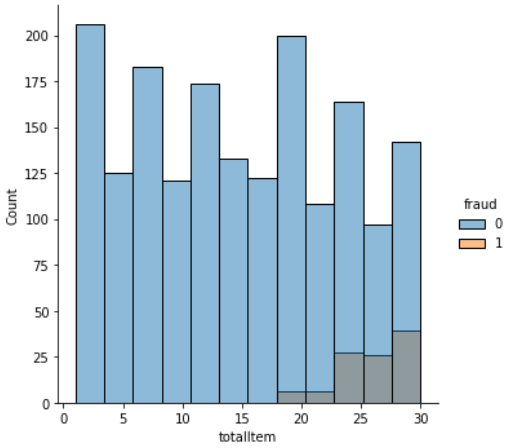
Feature Engineering

Menambahkan 1 fitur

```
totalItem = totalScanTimeInSeconds x scannedLineItemsPerSecond
```

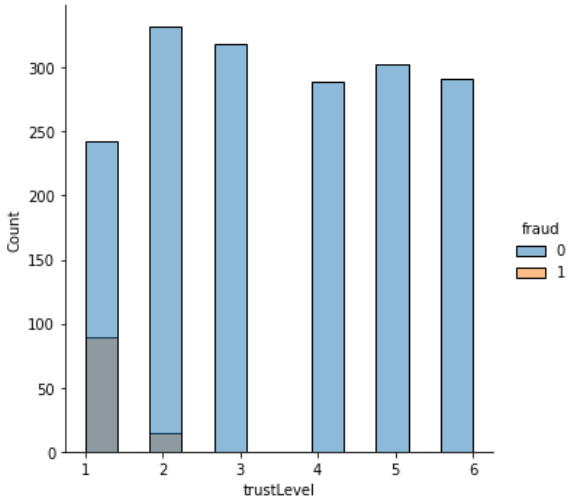
Fitur ini ditambahkan karena memiliki korelasi yang tinggi terhadap *fraud*

```
fraud                          1.000000
trustLevel                     0.319765
totalItem                      0.298423
totalScanTimeInSeconds         0.110414
lineItemVoidsPerPosition       0.090116
scansWithoutRegistration       0.074123
lineItemVoids                  0.063496
valuePerSecond                 0.028873
scannedLineItemsPerSecond      0.023085
grandTotal                     0.001421
quantityModifications          0.000864
```



Bisa dibilang untuk totalItem < 20 merupakan transaksi tidak *fraud*, model pun akan lebih optimal berlatih pada data *train*

Terlihat korelasi trustLevel paling tinggi yaitu 0.319765, sekarang kita lihat distribusinya



Sama seperti sebelumnya, untuk trustLevel > 2 merupakan transaksi tidak *fraud,* dengan demikian model bisa lebih optimal lagi berlatih pada data *train.*

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

Rancangan Sistem

Untuk melatih model, kita harus membagi data train menjadi X sebagai fitur dan y sebagai label, tidak hanya itu, X dan y kita bagi lagi menjadi data *train* dan data *validation*

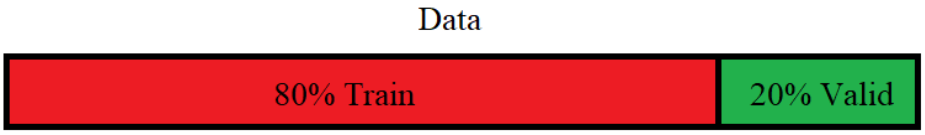
a. Data train

   Data yang digunakan untuk melatih model *machine learning*, model akan melihat dan belajar dari data ini.

b. Data Validation

   Data validation digunakan untuk mengevaluasi model, data ini tidak digunakan model untuk belajar, namun data ini digunakan untuk memperbarui parameter pada model agar lebih optimal

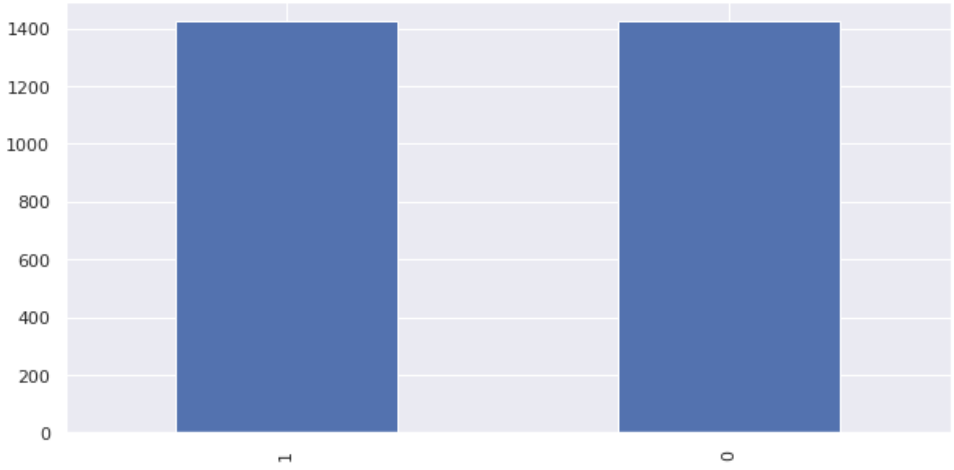
Pembagian data *train* dan data *validation* porsinya sebagai berikut:



Data yang diberikan sangat tidak *balance* atau *imbalance*. Data *imbalance* sangat mempengaruhi model belajar, data ini akan mempengaruhi kinerja model yang akan dibuat. Klasifikasi yang dilakukan akan membuat model belajar secara bias, sehingga akurasi pada kelas minoritas akan lebih kecil dibanding kelas mayoritas. Oleh karena itu, akan dilakukan *oversampling* pada data *train*. *Oversampling* adalah metode untuk menduplikasi kelas minoritas sampai berukuran sama dengan atau mendekati ukuran kelas mayoritas.

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

Metode *oversampling* yang digunakan adalah Synthetic Minority Oversampling Technique (SMOTE). Cara kerjanya yaitu dengan memilih salah satu data pada kelas minoritas, kemudian k dari tetangga terdekat dipilih secara acak menjadi *sample* baru. Prosedur ini terus berulang sampai kelas minoritas berukuran sama dengan kelas mayoritas.



Setelah itu kita inisialisasi model klasifikasi apa saja yang akan kita coba, di sini penulis menggunakan.

1. K-Nearest Neighbors
2. Support Vector Machine (SVM)
3. Linear SVC
4. Decision Tree
5. Random Forest
6. AdaBoost
7. GradientBoost
8. Naïve Bayes
9. Balance Bagging
10. RUSBoost
11. Linear Discriminant Analysis
12. Quadratic Discriminant Analysis

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

Dari beberapa model klasifikasi tersebut, kita coba latih satu persatu ke data train, dan dapatkan laporan klasifikasi dari prediksi data *validation*.

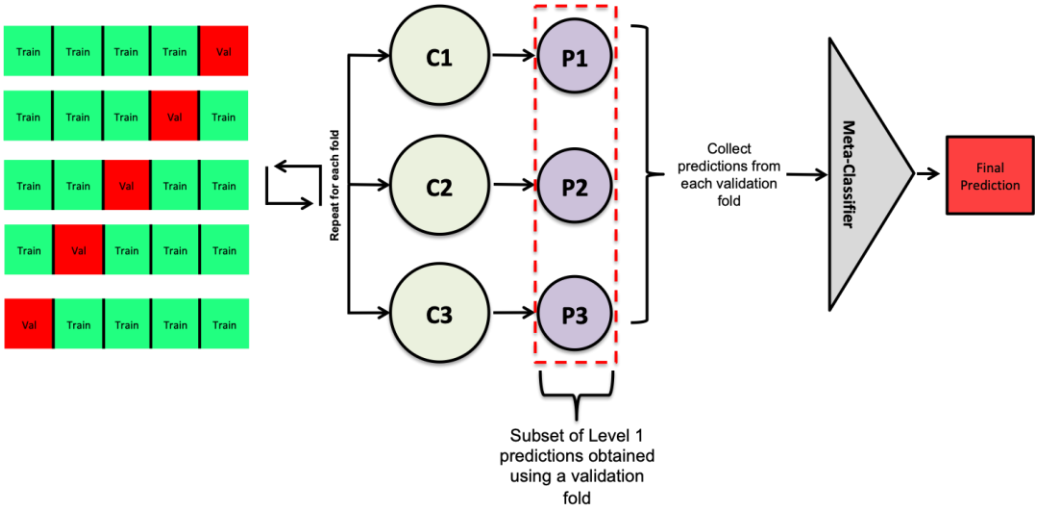
Hasil model-model kelasifikasi, sebagai berikut

Classifier	F1 Score (fraud)		Akurasi
	Kelas 0	Kelas 1	
KNN	0.85	0.17	74.7340 %
SVM	1.00	0.98	**99.7340 %**
Linear SVC	0.86	0.35	77.6595 %
Decision Tree	0.99	0.84	97.8723 %
Random Forest	0.99	0.86	98.1383 %
AdaBoost	1.00	0.96	**99.4681 %**
GradientBoost	0.99	0.90	**98.6702 %**
Naïve Bayes	0.95	0.56	90.4255 %
Balance Bagging	0.99	0.89	98.6702%
RUSBoost	0.99	0.83	97.6064%
Linear DA	0.93	0.51	87.7660 %
Quadratic DA	0.97	0.69	94.4149 %

Dengan demikian, kita mendapat 3 metode kelasifikasi terbaik, yaitu :

1. SVM
2. AdaBoost
3. GradientBoost

Selanjutnya kita gabungkan ketiga model tersebut menggunakan Stacking Classifier



Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

Model Stacking Classifier kemudian dilatih dengan data *train* dan diuji dengan data *validation*. Hasil laporan klasifikasi dari model terhadap data *validation* sebagai berikut:

Classifier	F1 Score (fraud)		Akurasi
	Kelas 0	Kelas 1	
Stacking	1.00	1.00	100 %

```
Accuracy: 100.0000%
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       352
           1       1.00      1.00      1.00        24

    accuracy                           1.00       376
   macro avg       1.00      1.00      1.00       376
weighted avg       1.00      1.00      1.00       376
```

Hasilnya cukup memuaskan, yaitu akurasi = 100%.

Evaluasi Model

Kemudian, dilakukan prediksi pada data *test* menggunakan model Stacking Classifier yang sudah dilatih sebelumnya.

Dikarenakan saat penulis membuat model, sudah terdapat solusi untuk data *test* ini. Maka dari itu, kita bisa mencari akurasi model terhadap data *test* ini.

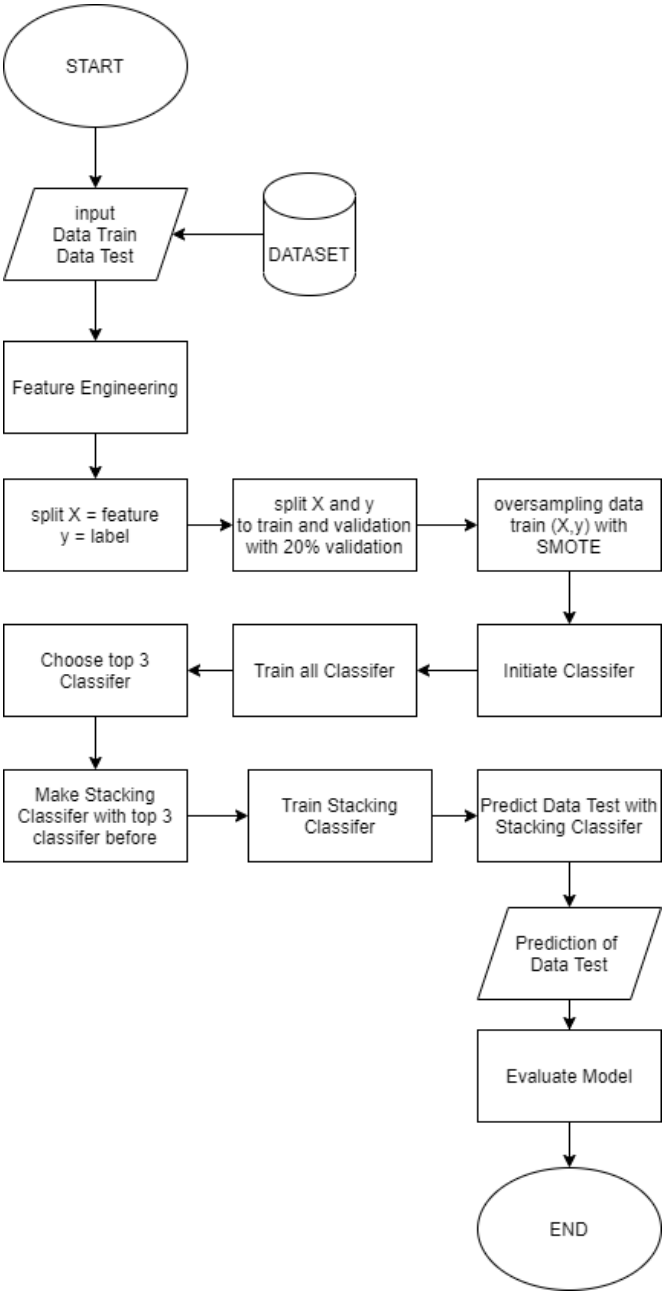
Setelah memprediksi data *test* dan dibandingkan dengan solusi yang ada, hasilnya sebagai berikut.

Classifier	F1 Score (fraud)		Akurasi
	Kelas 0	Kelas 1	
Stacking	1.00	0.91	99.0913 %

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00    474394
           1       0.89      0.92      0.91     23727

    accuracy                           0.99    498121
   macro avg       0.95      0.96      0.95    498121
weighted avg       0.99      0.99      0.99    498121

True positive =  471825
False positive =  2569
False negative =  1957
True negative =  21770
```

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224

FLOWCHART SISTEM



File notebook bisa dilihat di : https://github.com/ridaffa/dmc2019/blob/main/tubes.ipynb

Nama: Muhammad Ridaffa Purnomo
NIM: 1301174224