Annissa Claryta Berliana Putri (50421177)

CREDITA Aplikasi Pendeteksi Kelayakan Debitur

Tahukah anda?

Pengajuan kredit oleh debitur meningkat



Hal ini bisa menjadi **peluang** bagi pihak yang memiliki **usaha kredit** untuk meningkatkan pendapatan

Masih banyak orang mengajukan kredit tetapi tidak mampu membayar tagihan

Masalah



Hal ini dapat **menurunkan** pendapatan kreditur



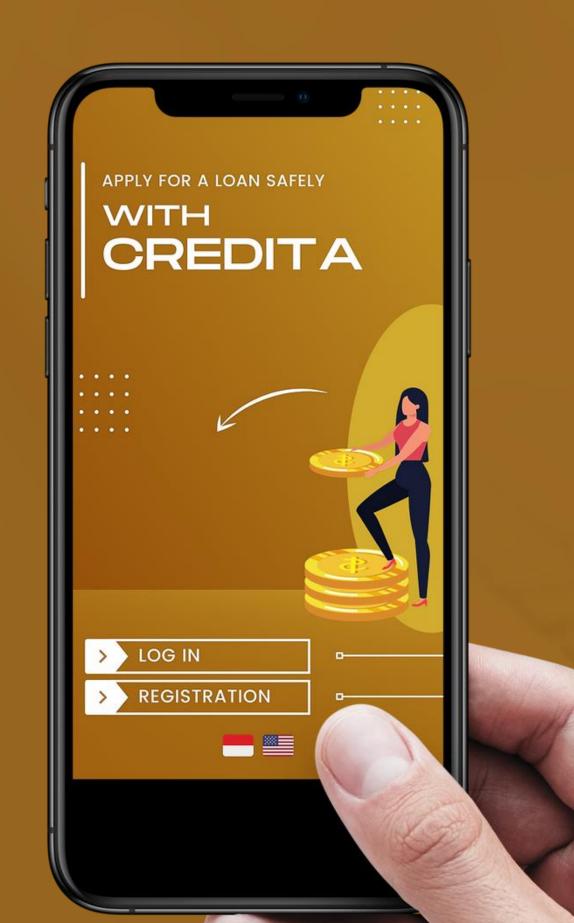
menyebabkan perusahaan mengalami **kerugian**



dan pada akhirnya **bangkrut**

Tantangan

Bagaimana untuk mengoptimalkan penerimaan pengajuan kredit?



Dengan melakukan **pengolahan data debitur** yang telah melakukan pinjaman...

Aplikasi CREDITA

Pendeteksi Kelayakan Pengajuan Kredit oleh Debitur dengan menggunakan Teknologi

Machine Learning

Implementasi **Machine Learning** dalam Bidang **Keuangan**

Metode

Teknologi **Machine Learning** adalah mesin yang dikembangkan untuk bisa **belajar dengan sendirinya** tanpa arahan dari penggunanya

Perkembangan ML **sangat pesat** dan penggunaannya dapat diimplementasikan dalam **berbagai hal**





Dengan membandingkan tiga model...



Logistic Regression



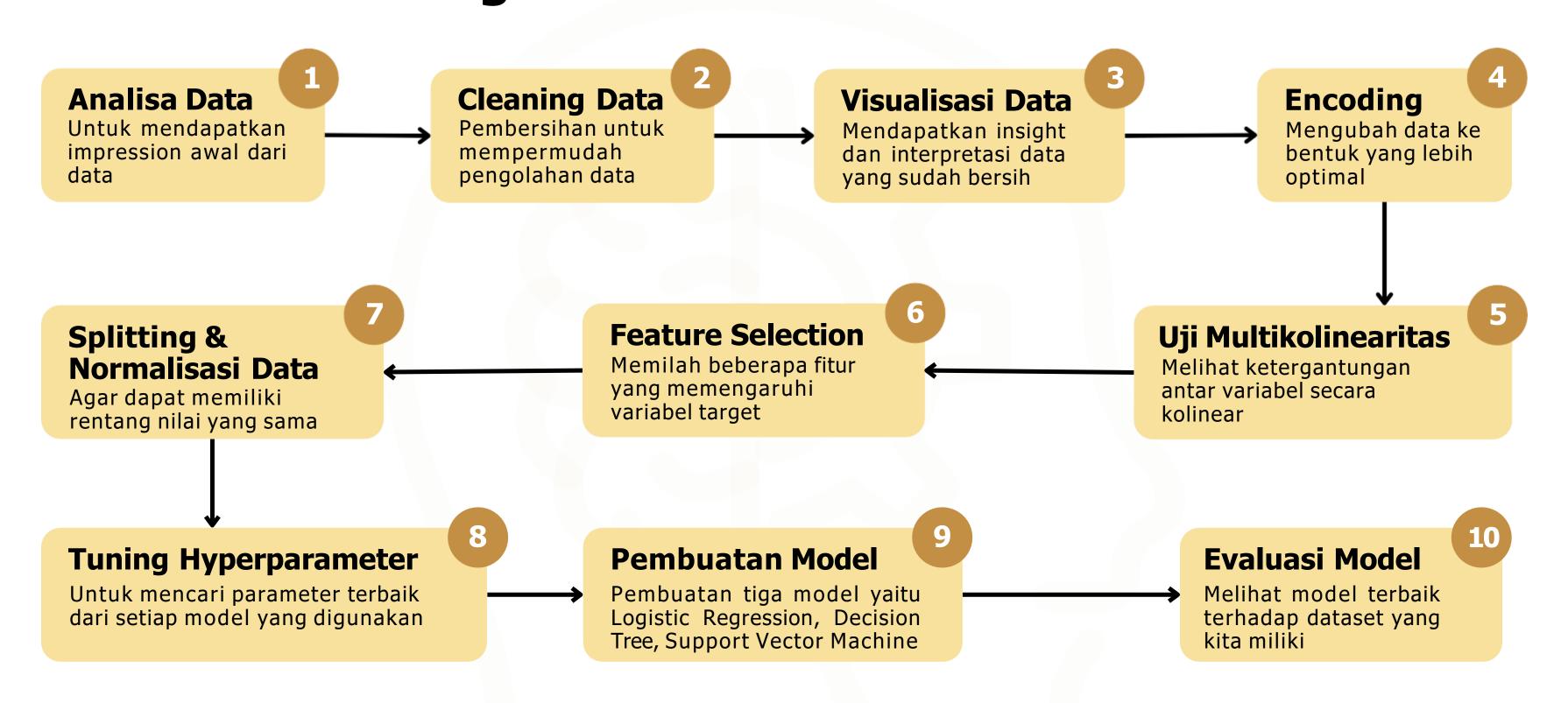
Decision Tree



dan akan dicari model terbaik untuk prediksi yang lebih akurat dan tepercaya

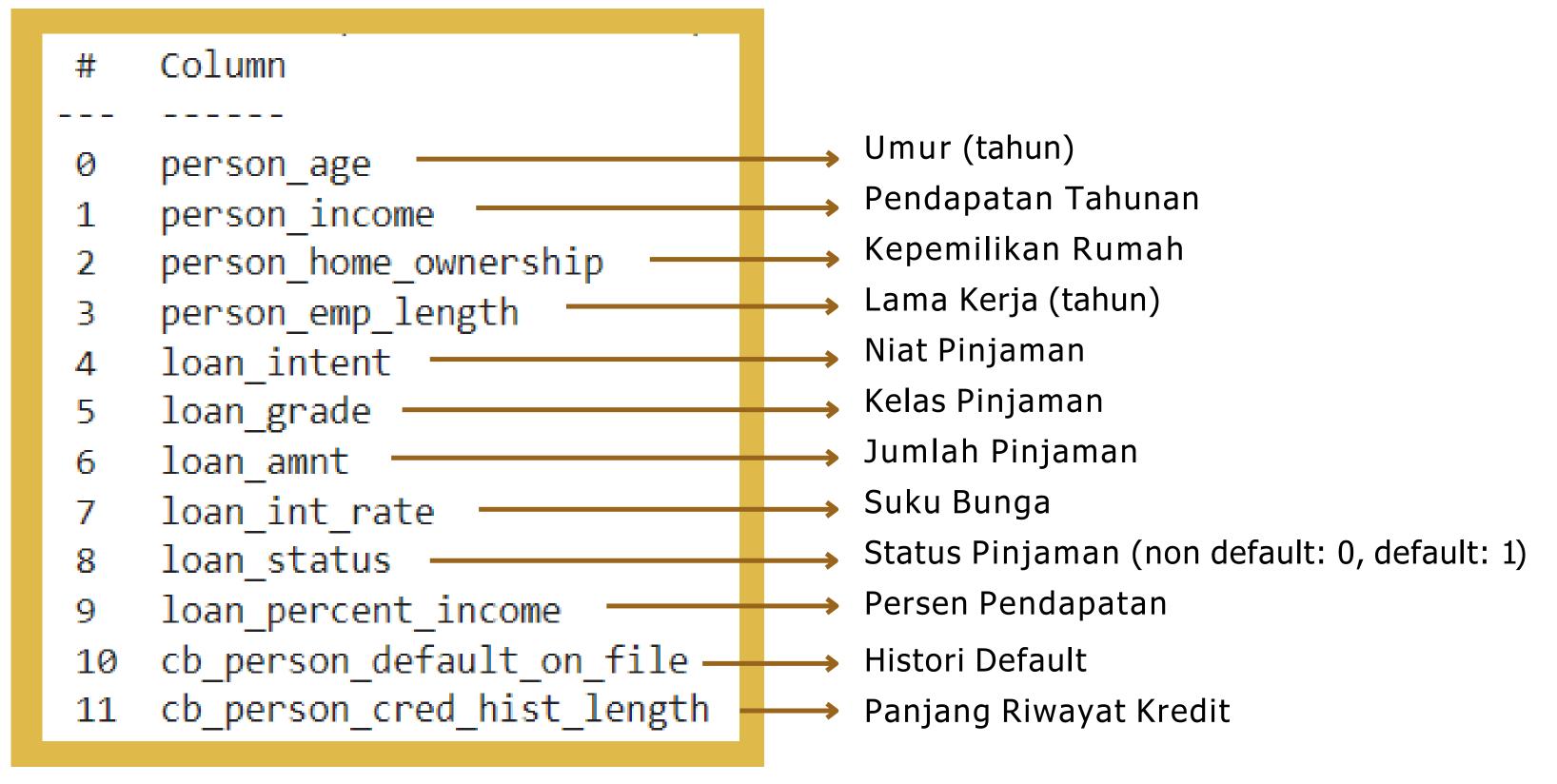
Langkah-langkah Pembuatan Model pada Machine Learning

Metode



Data debitur pada Dataset

Analisis Data



Ukuran data: 32581 baris x 12 kolom

Variabel Dependent dan Independent

Analisis Data

per	son_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate 1	oan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_lengt
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1	0.59	Y	
1	21	9600	OWN	5.0	EDUCATION	В	1000	11.14	0	0.10	N	
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1	0.57	N	
3	23	65500	RENT	4.0	MEDICAL	С	35000	15.23	1	0.53	N	3
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1	0.55	Υ	

Variabel Target (Dependent)

Keterangan Variabel Target: Ioan_status

- 1 (non-default) maka debitur tersebut tidak berisiko
- 2(default) menandakan debitur tersebut **berisiko**



Kolom lainnya akan menjadi variabel bebas (Independent)

Karakteristik Statistik Data

Analisis Data

Menunjukkan rangkuman statistik seperti rata-rata, median, dan data statistik lainnya

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	32581.000000	3.258100e+04	31686.000000	32581.000000	29465.000000	32581.000000	32581.000000	32581.000000
mean	27.734600	6.607485e+04	4.789686	9589.371106	11.011695	0.218164	0.170203	5.804211
std	6.348078	6.198312e+04	4.142630	6322.086646	3.240459	0.413006	0.106782	4.055001
min	20.000000	4.000000e+03	0.000000	500.000000	5.420000	0.000000	0.000000	2.000000
25%	23.000000	3.850000e+04	2.000000	5000.000000	7.900000	0.000000	0.090000	3.000000
50%	26.000000	5.500000e+04	4.000000	8000.000000	10.990000	0.000000	0.150000	4.000000
75%	30.000000	7.920000e+04	7.000000	12200.000000	13.470000	0.000000	0.230000	8.000000
max	144.000000	6.000000e+06	123.000000	35000.000000	23.220000	1.000000	0.830000	30.000000

terdapat kemungkinan **outliers** dimana...

Data Maksimum dari person_age sebesar 144 tahun

Data Maksimum dari person_emp_length sebesar 123 tahun



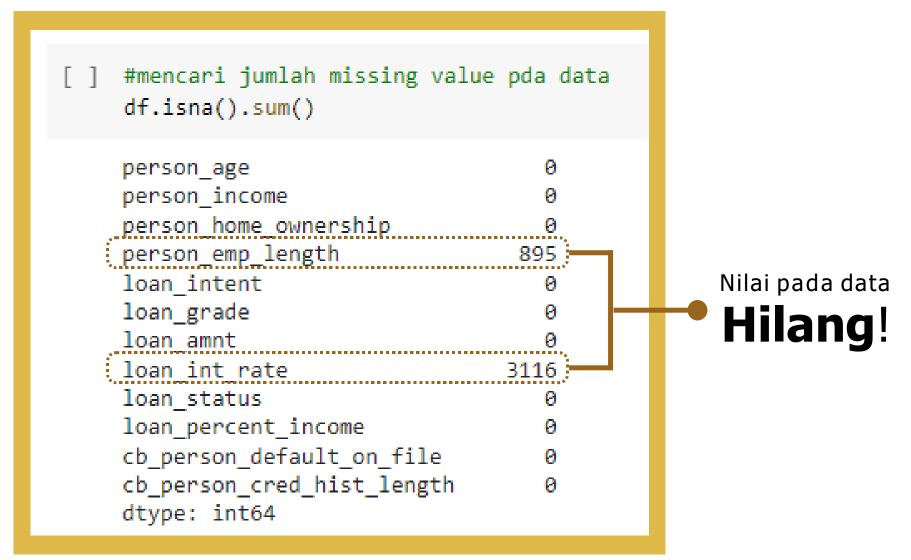


Mendeteksi Missing Value

Cleaning Data

Missing value merupakan nilai atribut yang kosong pada objek data

Q Pencarian



Terdapat 2 kolom yang mengandung missing value

person_emp_length & loan_int_rate

Penanganan

Akan dilakukan penanganan dengan imputasi data

Missing value akan diganti dengan nilai rata-rata data

```
#mengubah missing value person_emp_length menjadi rata-rata
persempmean = round(df['person_emp_length'].mean())
df['person_emp_length'] = df['person_emp_length'].fillna(persempmean)

#mengubah missing value loan_int_rate menjadi rata-rata
loanmean = df['loan_int_rate'].mean()
df['loan_int_rate'] = df['loan_int_rate'].fillna(loanmean)
```

Ketika dilakukan pengecekan missing value kembali...

```
person_age0loan_amnt0person_income0loan_int_rate0person_home_ownership0loan_status0person_emp_length0loan_percent_income0loan_intent0cb_person_default_on_file0loan_grade0cb_person_cred_hist_length0
```

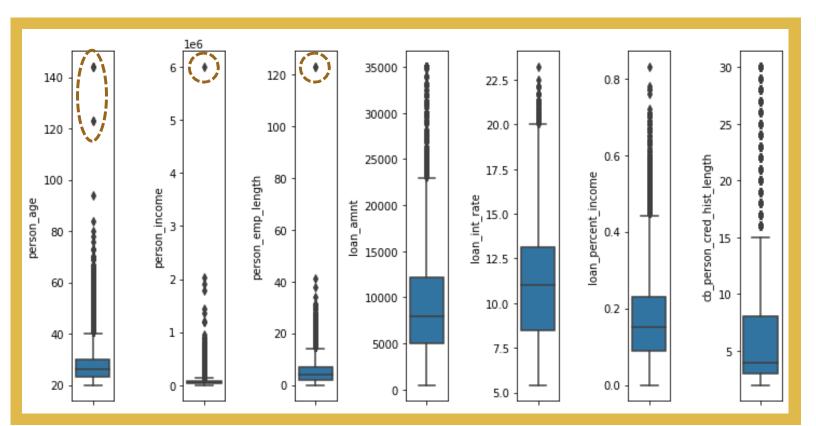
Sudah tidak ada data kosong

Mendeteksi Outliers

Cleaning Data

Outlier adalah data observasi yang muncul dengan nilai-nilai ekstrim

Q Pencarian



Terdeteksi Nilai Ekstrim!

Terdapat **3 kolom** yang mengandung outliers

person_age, person_income, persone_emp_length

Penanganan

Akan dilakukan penanganan dengan membuang outliers

```
#memeriksa data person_income yang menjadi outlier
df.loc[df['person_income'] == 6000000]
#menghapus row
df.drop([32297], axis=0, inplace=True)

df.loc[df['person_age'] >100]
#menghapus row
df.drop([81,183,575,747], axis=0, inplace=True)

df.loc[df['person_emp_length'] >df['person_age']]
#menghapus row
df.drop([0,210], axis=0, inplace=True)
```

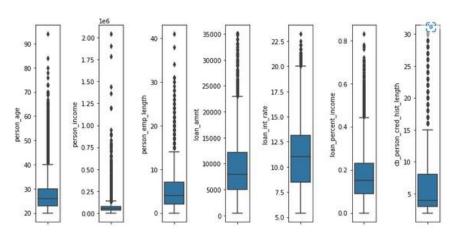
Drop data **person_income** sebesar 6000000

Drop data **person_age** yang lebih besar dari 100

Drop data **person_emp_length** yang lebih besar dari person_age

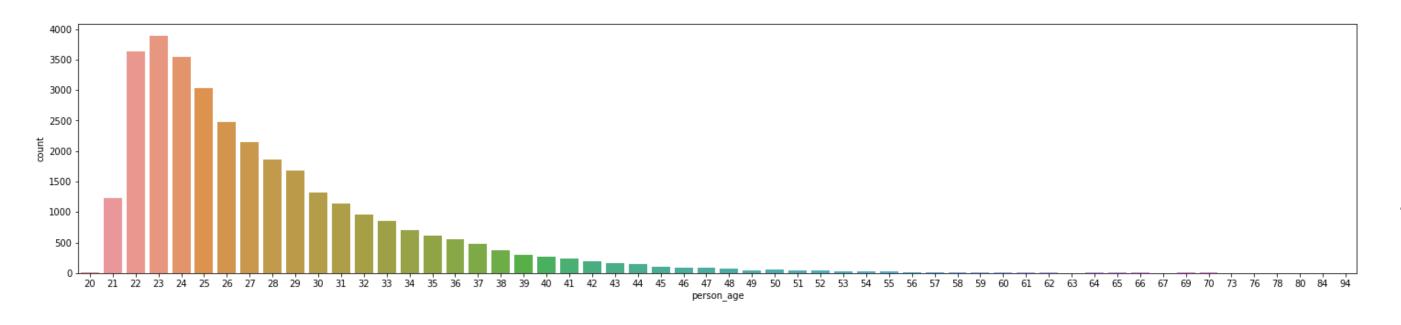
Ketika dilakukan pengecekan outliers kembali...

Sudah **tidak ada outliers**

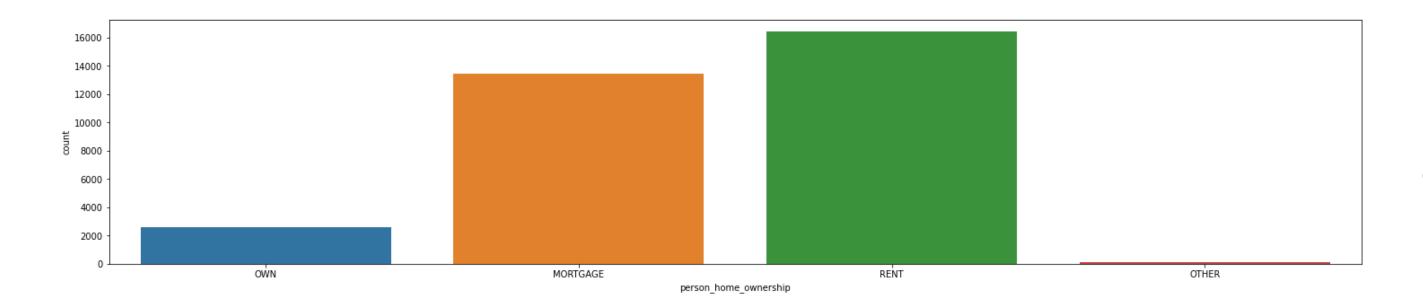


Visualisasi Data

Mengkonversikan kumpulan data menjadi hal yang lebih sederhana untuk disajikan



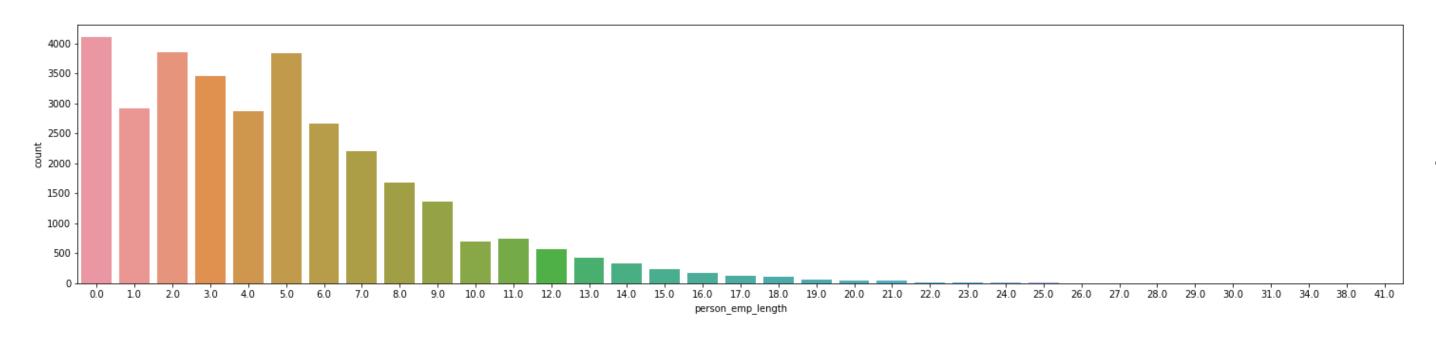
Sebagian besar debitur berusia produktif yakni pada rentang 21-40 tahun dan jumlah debitur terbanyak ada pada usia 23 tahun



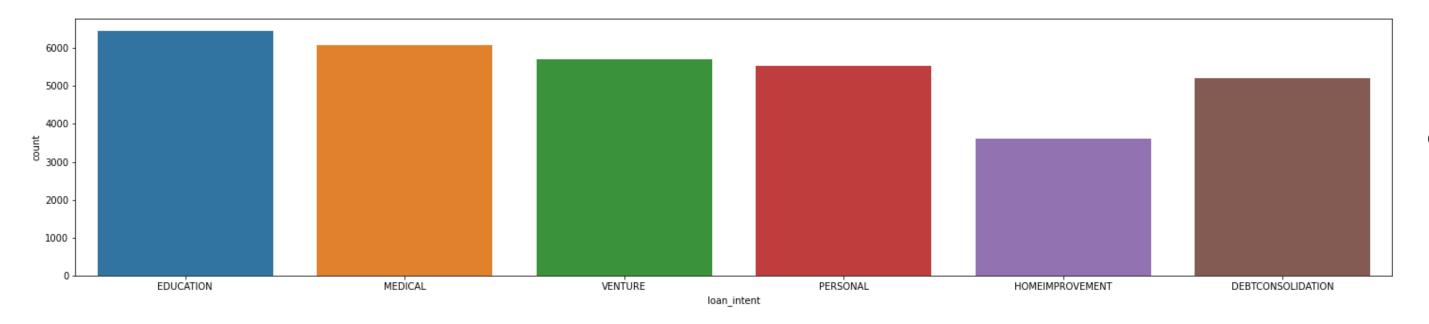
Urutan jenis kepemilikan rumah debitur dari tertinggi hingga terendah adalah rent (sewa), mortgage (gadai), own (milik pribadi), dan other (lain-lain)

Visualisasi Data

Mengkonversikan kumpulan data menjadi hal yang lebih sederhana untuk disajikan



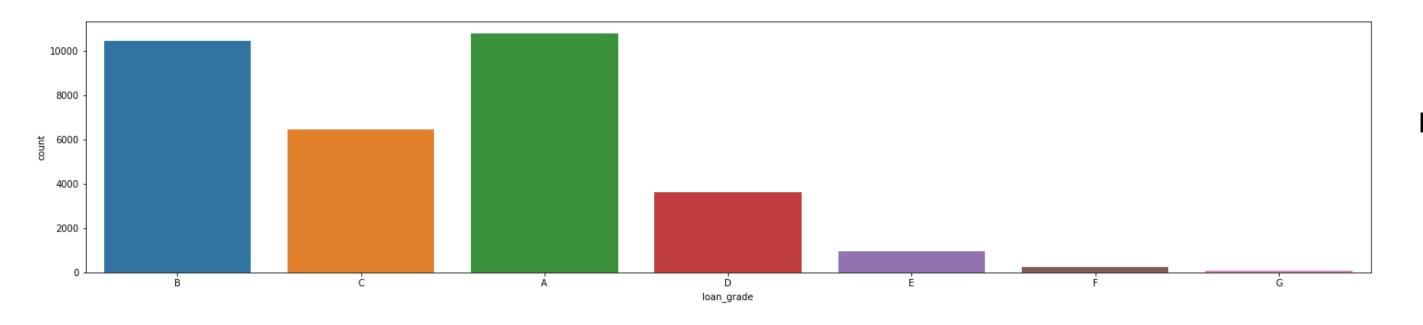
Debitur memiliki lama bekerja berkisar **0 hingga 20 tahun** dan debitur terbanyak memiliki lama kerja **0 tahun**



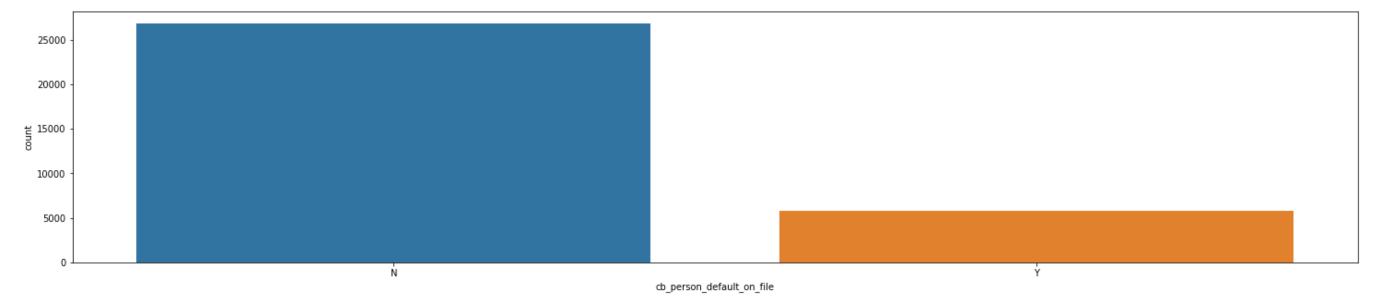
Mayoritas tujuan kredit adalah untuk kebutuhan edukasi, diikuti biaya medis, usaha, kebutuhan pribadi, pembayaran hutang, dan renovasi rumah

Visualisasi Data

Mengkonversikan kumpulan data menjadi hal yang lebih sederhana untuk disajikan



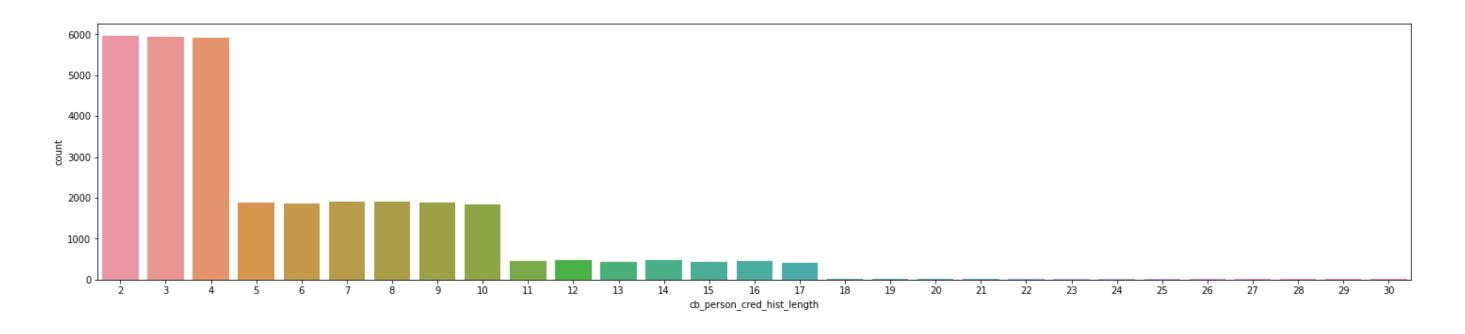
Grade tertinggi yang menjadi pilihan kreditur adalah **grade A** yang kemudian **disusul oleh grade B** dimana grade A memiliki risk yang paling kecil



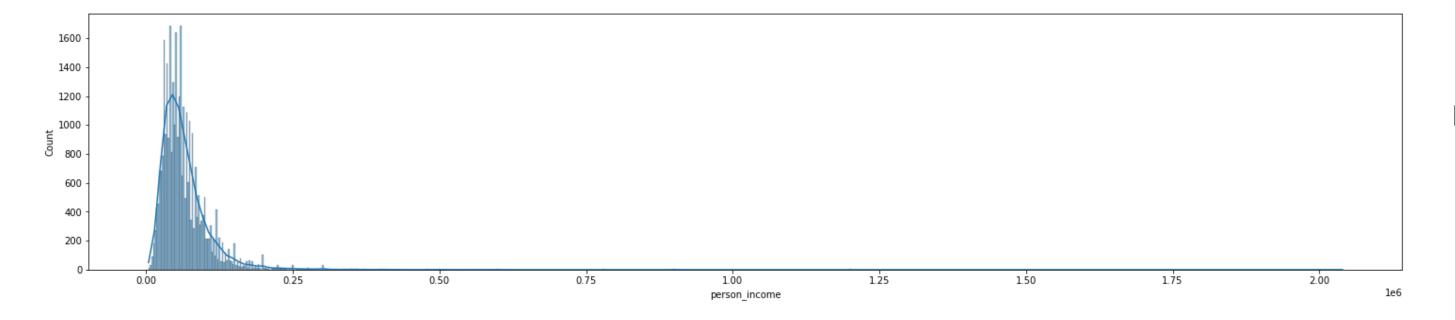
Lebih banyak debitur dengan riwayat **berhasil membayar pinjaman** dibanding dengan debitur yang pernah mengalami kegagalan pembayaran kredit

Visualisasi Data

Mengkonversikan kumpulan data menjadi hal yang lebih sederhana untuk disajikan



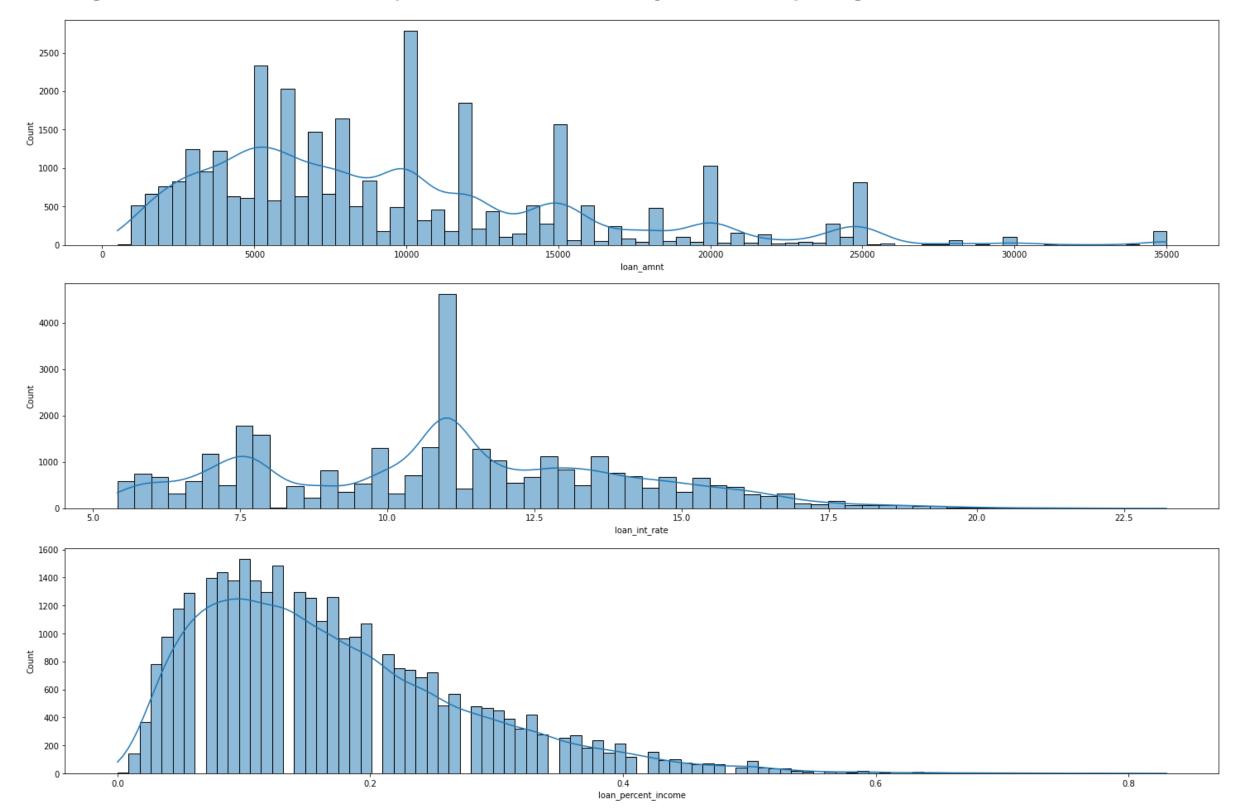
Mayoritas panjang riwayat kredit debitur berada di rentang waktu **2-4 tahun**



Pendapatan tahunan debitur berada pada rentang puluhan hingga ratusan ribu dan paling banyak berada di kisaran **60000**

Visualisasi Data

Mengkonversikan kumpulan data menjadi hal yang lebih sederhana untuk disajikan



Jumlah pinjaman tergolong bervariasi dengan pinjaman tertinggi berada di kisaran **10000**

Suku bunga pinjaman juga bervariasi berkisar dari **5.42% ke 23.2%** dengan suku bunga yang paling banyak dipilih adalah **11%**

Persenan pendapatan berada di rentang **0 hingga 0.83** dengan modus data dipegang oleh **0.1**

Melakukan Encoding

Mengubah fitur yang bertipe object menjadi numerik

Q Pencarian

df.select_dtypes(include=['object'])

Hasil:

perso	n_home_ownership	loan_intent	loan_grade	cb_person_default_on_file	
1	OWN	EDUCATION	В	N	
2	MORTGAGE	MEDICAL	С	N	
3	RENT	MEDICAL	С	N	
4	RENT	MEDICAL	C	Υ	
5	OWN	VENTURE	Α	N	
1222	222	99	222)	(a.c)	
32576	MORTGAGE	PERSONAL	С	N	
32577	MORTGAGE	PERSONAL	Α	N	
32578	RENT	HOMEIMPROVEMENT	В	N	
32579	MORTGAGE	PERSONAL	В	N	
32580	RENT	MEDICAL	В	N	

Terdapat 4 kolom yang bertipe object

person_home_ownership, loan_intent,
loan_grade, cb_person_default_on_file

Penanganan

Akan dilakukan penanganan dengan melakukan encoding

One Hot Encoding dilakukan untuk kolom-kolom bertipe nominal

```
# OneHot encoding
i = ['person_home_ownership', 'loan_intent', 'cb_person_default_on_file']
OneHot = pd.get_dummies(df, columns = i)
df = OneHot
```

Ordinal Encoding dilakukan untuk kolom-kolom bertipe ordinal

```
# Ordinal encoding
df.loan_grade = pd.Categorical(df.loan_grade)
df['loan_grade'] = df.loan_grade.cat.codes
```

Ketika dilakukan pengecekan tipe kembali...

Sudah tidak ada data yang bertipe object

```
Data columns (total 21 columns):
# Column
                                   Non-Null Count Dtype
                                   -----
    person age
                                   32574 non-null int64
                                   32574 non-null int64
    person_income
                                   32574 non-null int64
    person emp length
                                   32574 non-null int8
    loan_grade
    loan amnt
                                   32574 non-null int64
                                   32574 non-null float64
    loan int rate
                                   32574 non-null int64
    loan_status
    loan percent income
                                   32574 non-null float64
    cb_person_cred_hist_length
                                  32574 non-null int64
    person home ownership MORTGAGE 32574 non-null uint8
    person home ownership OTHER
                                  32574 non-null uint8
                                   32574 non-null uint8
 11 person_home_ownership_OWN
 12 person home ownership RENT
                                   32574 non-null uint8
 13 loan_intent_DEBTCONSOLIDATION
                                  32574 non-null uint8
 14 loan intent EDUCATION
                                   32574 non-null uint8
 15 loan intent HOMEIMPROVEMENT
                                  32574 non-null uint8
 16 loan_intent_MEDICAL
                                   32574 non-null uint8
                                   32574 non-null uint8
 17 loan intent PERSONAL
 18 loan_intent_VENTURE
                                   32574 non-null uint8
 19 cb_person_default_on_file_N
                                  32574 non-null uint8
 20 cb person default on file Y
                                  32574 non-null uint8
dtypes: float64(2), int64(6), int8(1), uint8(12)
memory usage: 2.6 MB
```

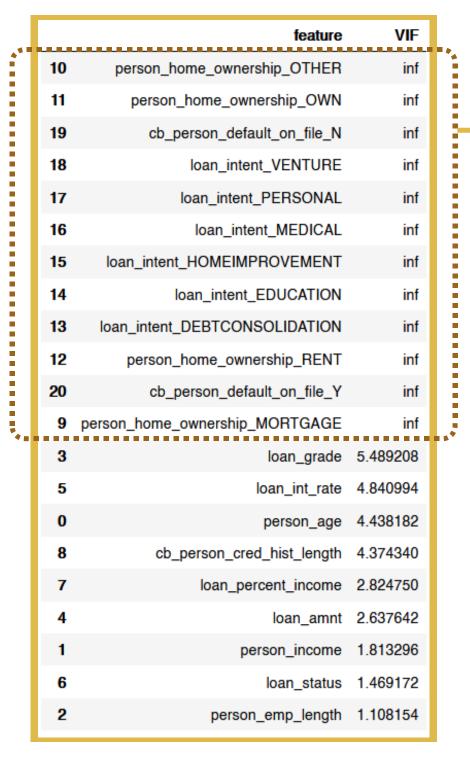
Encoding

Melakukan Uji Multikolinearitas

Uji Multikolinearitas

Untuk melihat bahwa tidak ada fitur (variabel independen) yang **saling berkolerasi satu sama lain** dalam sebuah model regresi

Q Dengan Variance Inflation Factor (VIF)



Penjelasan

VIF's Score = **inf**, mengartikan bahwa terdapat korelasi sempurna pada di antara variabel yang ada sehingga variabel tersebut **redundant**.

Akan tetapi, kasus ini terjadi karna fitur yang memiliki **VIF's Score = inf** merupakan fitur hasil dari OneHot Encoding, sehingga memiliki korelasi yang sempurna. Maka fitur-fitur tersebut tetap akan digunakan.

Untuk fitur lainnya memiliki nilai multikoliniearitas < 10, maka tidak terjadi multikolinearitas

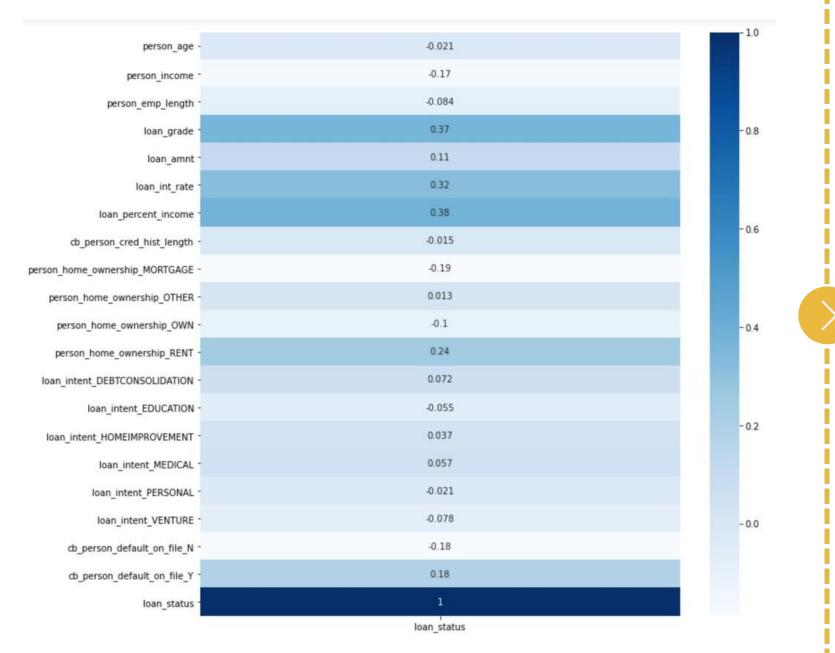
Kesimpulan

Kita dapat menggunakan **semua fitur** untuk tahap feature selection

Feature Selection dengan Pearson Correlation Feature Selection

Pearson Correlation merupakan metode yang digunakan untuk mencari hubungan antara independent variable dengan dependent variable.





† Penjelasan

Telah didapatkan korelasi antara fitur-fitur dengan variabel target yaitu 'loan status'. Kemudian akan dicari mean dari seluruh nilai korelasi dan variabel dengan korelasi yang lebih besar dari nilai mean akan dipilih.

```
Mean dari korelasi tiap variabel yang ada adalah 0.140147762993015
Sehingga, variabel yang dipilih sebagai hasil seleksi fitur adalah
person income
                                  0.168485
loan grade
                                  0.373042
loan_int_rate
                                  0.319302
loan_percent_income
                                  0.379250
person_home_ownership_MORTGAGE
                                  0.187607
person_home_ownership_RENT
                                  0.238416
cb person default on file N
                                  0.179002
cb person default on file Y
                                  0.179002
```

Splitting dan Normalisasi Data

Splitting dan Normalisasi Data

Q Hasil

† Penjelasan

X adalah dataframe berisi fitur-fitur yang terpilih dari proses feature selection

Y adalah variabel target

```
# train - test split dengan perbandingan 8 : 2
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 = 0)
```

Membagi data train dan data tes dengan perbandingan 8:2

26064

6516

```
# Normalisasi dengan MinMaxScaler
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler()
X_train = sc.fit_transform(X_train) # fit X_train
X_test = sc.transform(X_test) #transform X_test
```

Min Max Scaler dilakukan untuk mengubah data numerik menjadi range [0,1]

Melakukan Tuning Hyperparameter

Tuning Hyperparameter

Untuk mencari parameter terbaik dari setiap model yang digunakan

Q GridSearchCV

GridSearchCV adalah teknik untuk mencari **nilai parameter terbaik** dari kumpulan parameter grid yang diberikan

```
# scoring model and parameters, pencarian model dan parameter terbaik dengan menggunakan GridSearchCV
from sklearn.model_selection import GridSearchCV
for model_name, mp in model_params.items():
    clf = GridSearchCV(mp['model'], mp['params'], cv=3, return_train_score=False)
    clf.fit(X_train,y_train) # fit GridSearch ke X_train dan y_train
    scores.append({
        'model': model_name,
        'best_score': clf.best_score_,
        'best_params': clf.best_params_
})
best = pd.DataFrame(scores,columns=['model','best_score','best_params'])
```

	model	best_score	best_params
0	Logistic_Regression	0.850301	{'C': 0.1, 'multi_class': 'multinomial', 'pena
1	decision_tree	0.892590	{'criterion': 'gini', 'max_depth': 8, 'max_fea
2	SVM_Classifier	0.872558	{'gamma': 'scale', 'kernel': 'poly'}

Hasil

Model Logistic Regression

Best Score : 0.850301

Best Parameter:

'C': 0.1, 'multi_class': 'multinomial', 'penalty': 'none', 'solver': 'sag'

Model Decision Tree

Best Score: 0.892590

Best Parameter :

'criterion': 'gini', 'max_depth': 8, 'max_features': 'log2', 'splitter':

'best'

Model Support Vector Machine

Best Score: 0.872558

Best Parameter:

'gamma': 'scale', 'kernel': 'poly'

Logistic Regression

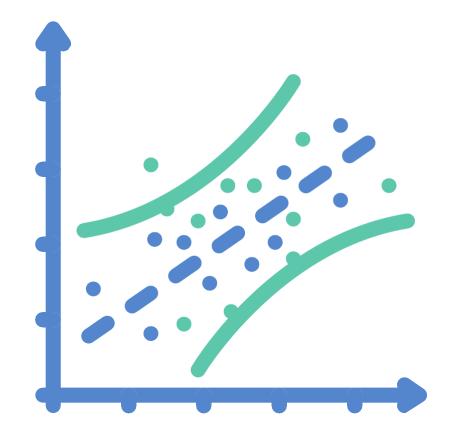
Pembuatan Model untuk Klasifikasi

Pembuatan Model



Metode

- Regresi Logistrik menggunakan konsep regresi untuk memprediksi suatu nilai yang bersifat kategorik
- Regresi Logistik memberikan output berupa
 probabilitas suatu item termasuk kelas tertentu





```
# Membentuk Model Logistic Regression

from sklearn.linear_model import LogisticRegression

LR = LogisticRegression(C = 0.1, solver = 'sag', multi_class = 'ovr', penalty = 'none').fit(X_train, y_train)
```

Decision Tree

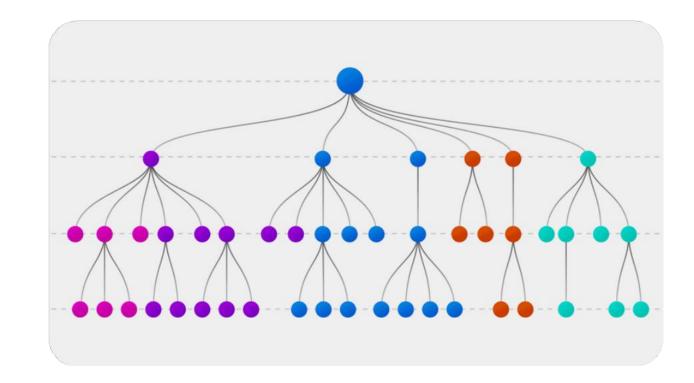
Pembuatan Model

Pembuatan Model untuk **Klasifikasi**



Metode

 Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki



```
Coding
```

```
# Membentuk Model Decision Tree
from sklearn.tree import DecisionTreeClassifier
credittree = DecisionTreeClassifier(criterion = 'entropy', max_depth = 8, random_state = 0, splitter = 'best')
```

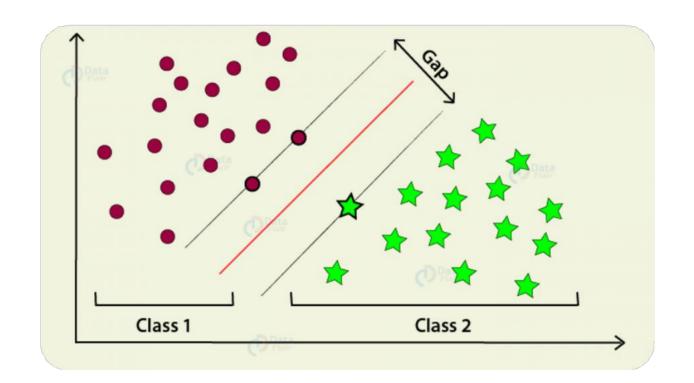
Support Vector Machine (SVM)

Pembuatan Model untuk Klasifikasi

Pembuatan Model



 SVM adalah algoritma supervised learning untuk klasifikasi dengan cara menemukan separator berupa hyperplane



Coding

```
# Membentuk Model Support Vector Machines (SVM)
from sklearn import svm
clf = svm.SVC(C=1, gamma = 'scale', kernel='poly')
```

Evaluasi **Model**

Melihat Model terbaik pada dataset

Evaluasi Model

support

0.88

0.79

0.87

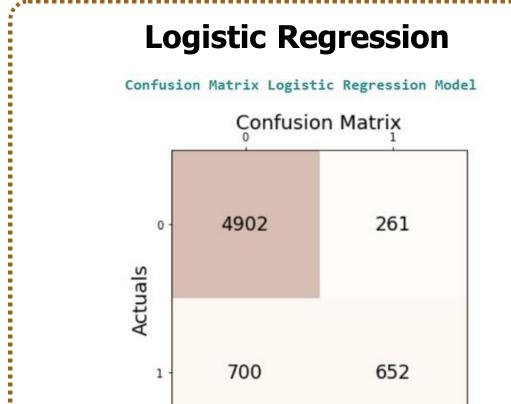
5163

1352

6515

6515

6515



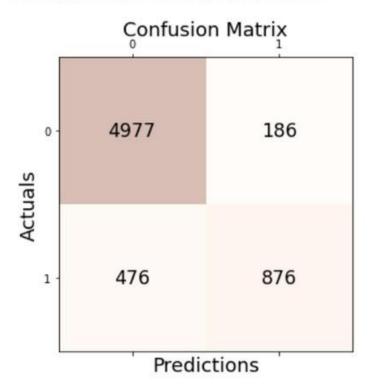
Classification Report Logistic Regression Model

	precision	recall	f1-score	support
0	0.88	0.95	0.91	5163
1	0.71	0.48	0.58	1352
accuracy			0.85	6515
macro avg	0.79	0.72	0.74	6515
weighted avg	0.84	0.85	0.84	6515

Predictions

f1-score: 0.9107292150487692

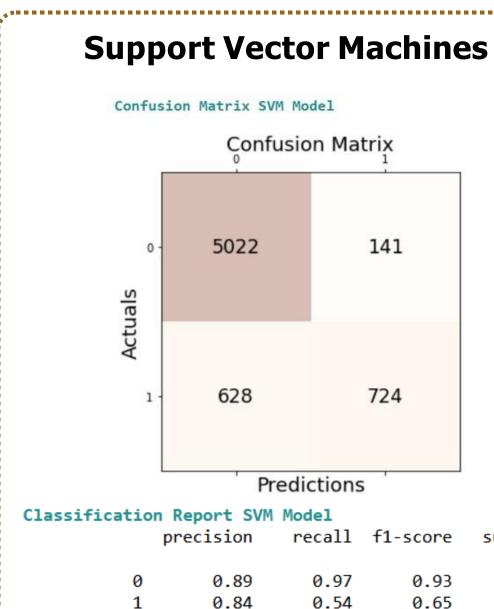
Decision Tree Confusion Matrix Decision Tree Model



Classification Report Decision Tree Model

	precision	recall	f1-score	support
0	0.91	0.96	0.94	5163
1	0.82	0.65	0.73	1352
accuracy macro avg weighted avg	0.87 0.89	0.81 0.90	0.90 0.83 0.89	6515 6515 6515

f1-score: 0.9376412961567445



0.86

0.88

f1-score: 0.9288819014149634

0.75

0.88

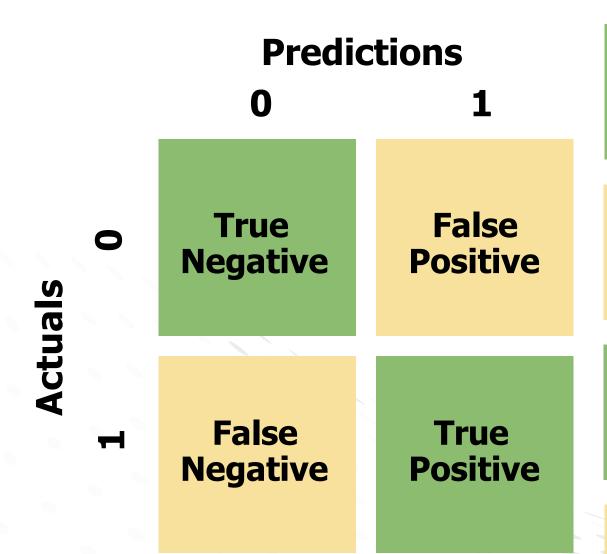
accuracy

macro avg

weighted avg

Evaluasi Model - Confusion Matrix

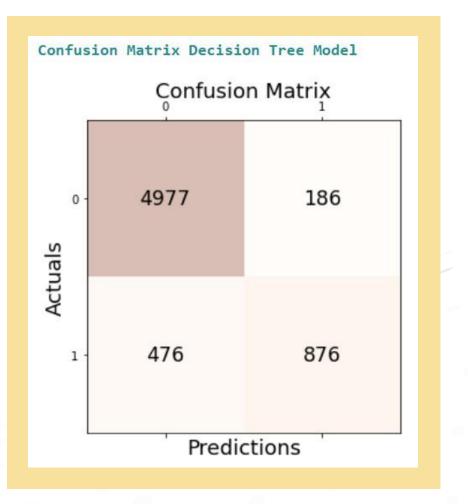
Melihat Model terbaik pada dataset



- 1) True Negative: Model memprediksi data ada di kelas **Negatif** dan yang sebenarnya data memang ada di kelas **Negatif**.
- 2) False Negative: Model memprediksi data ada di kelas **Negatif** dan yang sebenarnya data memang ada di kelas **Positif**.
- 3) True Positive: Model memprediksi data ada di kelas **Positif** dan yang sebenarnya data memang ada di kelas **Positif**.
- 4) False Positive: Model memprediksi data ada di kelas **Positif** dan yang sebenarnya data memang ada di kelas Negatif.

Evaluasi Model

Decision Tree



Evaluasi Model - Classification Report

Evaluasi Model

Melihat **model terba**ik pada dataset

Penjelasan

Decision Tree

Classification Report Decision Tree Model								
	precision	recall	f1-score	support				
0	0.91	0.96	0.94	5163				
1	0.82	0.65	0.73	1352				
accuracy			0.90	6515				
macro avg	0.87	0.81	0.83	6515				
weighted avg	0.89	0.90	0.89	6515				
f1-scope: 0	02764120615	67/1/15						
TI-Score: 0.	f1-score: 0.9376412961567445							

Precision: Perbandingan antara True Positive dengan banyaknya data yang diprediksi positif

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

Recall: Perbandingan antara True Positive dengan banyaknya data yang sebenarnya positif

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

F1-score: Harmonic mean dari precision dan recall

$$\frac{1}{F1} = \frac{1}{2} \left(\frac{1}{precision} + \frac{1}{recall} \right)$$

Support: Jumlah data test yang digunakan

Model terbaik yang kita peroleh menggunakan

Decision Tree

parameter 'criterion': 'gini', 'max_depth': 8, 'max_features': 'log2', 'splitter': 'best'

Interpretasi dari hasil yang kami peroleh memiliki kesimpulan bahwa untuk mengetahui debitur layak menerima kredit atau tidak dengan memperhatikan

- 1 Jumlah Pendapatan Tahunan
- **2** Kelas Pinjaman
- **3** Suku Bunga

- **4** Persen Pendapatan
- **5** Kepemilikan Rumah Gadai/Sewa
- 6 Riwayat Gagal Bayar

Python **Programming**

Credit Risk Prediction Programming Algorithm

Person Income: 9600

Loan Grade: c

Loan Interest Rate: 12.87 Loan Percent Income: 0.57

Person Home Ownership: mortGage CB Person Default on File: n

Kelayakan Penerima Nasabah Kredit:

Tidak Layak

Credit Risk Prediction Programming Algorithm

Person Income: 120000

Loan Grade: A

Loan Interest Rate: 7.49 Loan Percent Income: 0.15

Person Home Ownership: mortgage CB Person Default on File: n

Kelayakan Penerima Nasabah Kredit:

Layak

Credit Risk Prediction Programming Algorithm

Person Income: 76000

Loan Grade: B

Loan Interest Rate: 10.99
Loan Percent Income: 0.46
Person Home Ownership: RENT
CB Person Default on File: N

Kelayakan Penerima Nasabah Kredit:

Tidak Layak





••••

Implementasi



Mata Kuliah Sains Data – SCMA602017

TERIMA KASIH

Program Pendeteksi Kelayakan Debitur