



MODEL PREDIKSI RISIKO KREDIT

id/x partners

Rakamin
Academy

OKTOBER 2022

Final Project Data Scientist Virtual Intern

AGENDA

The objectives
Business Understanding
Data Preparation
Exploratory Data Analysis
Data Preprocessing
Modeling
Summary

**TODAY'S
HIGHLIGHTS**

APA ITU CREDIT RISK?

Risiko kredit adalah suatu risiko kerugian yang dialami perusahaan yang disebabkan oleh ketidakmampuan dari debitur atas kewajiban pembayaran utangnya. Salah satu faktor penyebab terjadinya risiko kredit adalah kesalahan penilaian dalam keputusan pemberian pinjaman.



BACKGROUND



BAGAIMANA MENGATASINYA?

Dengan membangun suatu model prediksi tingkat risiko kredit berdasarkan data historis atas transaksi yang pernah dilakukan oleh peminjam untuk melihat pola peminjam pada pembayaran kredit. Dalam hal ini akan digunakan algoritma machine learning.

BUSINESS UNDERSTANDING

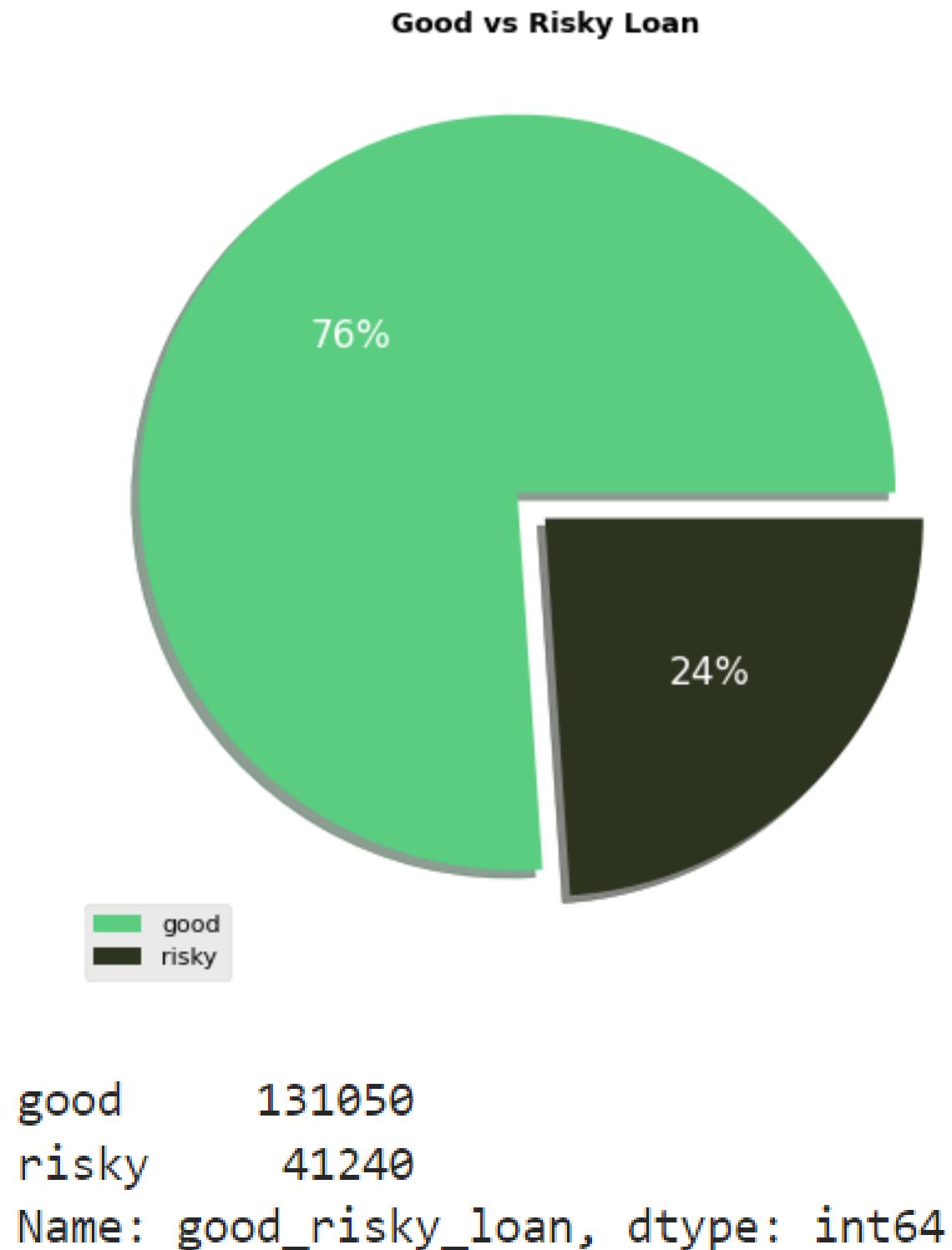
Tujuan dari project ini adalah untuk membangun model yang dapat memprediksi risiko kredit berdasarkan data pinjaman dari tahun 2007 hingga 2014.

Dari model tersebut dapat meminimalkan tingkat risiko kredit yang akan terjadi karena dapat mengklasifikasi peminjam mana yang pengajuan kreditnya perlu ditolak ataupun diterima.



DATA PREPARATION

- Data yang digunakan adalah data pinjaman dari tahun 2007 sampai 2014
- Target label pada model ini dibagi menjadi dua klasifikasi, yaitu **baik** (good) dan **berisiko** (risky)
- Peminjam diklasifikasikan baik, jika riwayat kreditnya fully paid, dan diklasifikasikan berisiko jika pernah memiliki riwayat telat ataupun gagal bayar

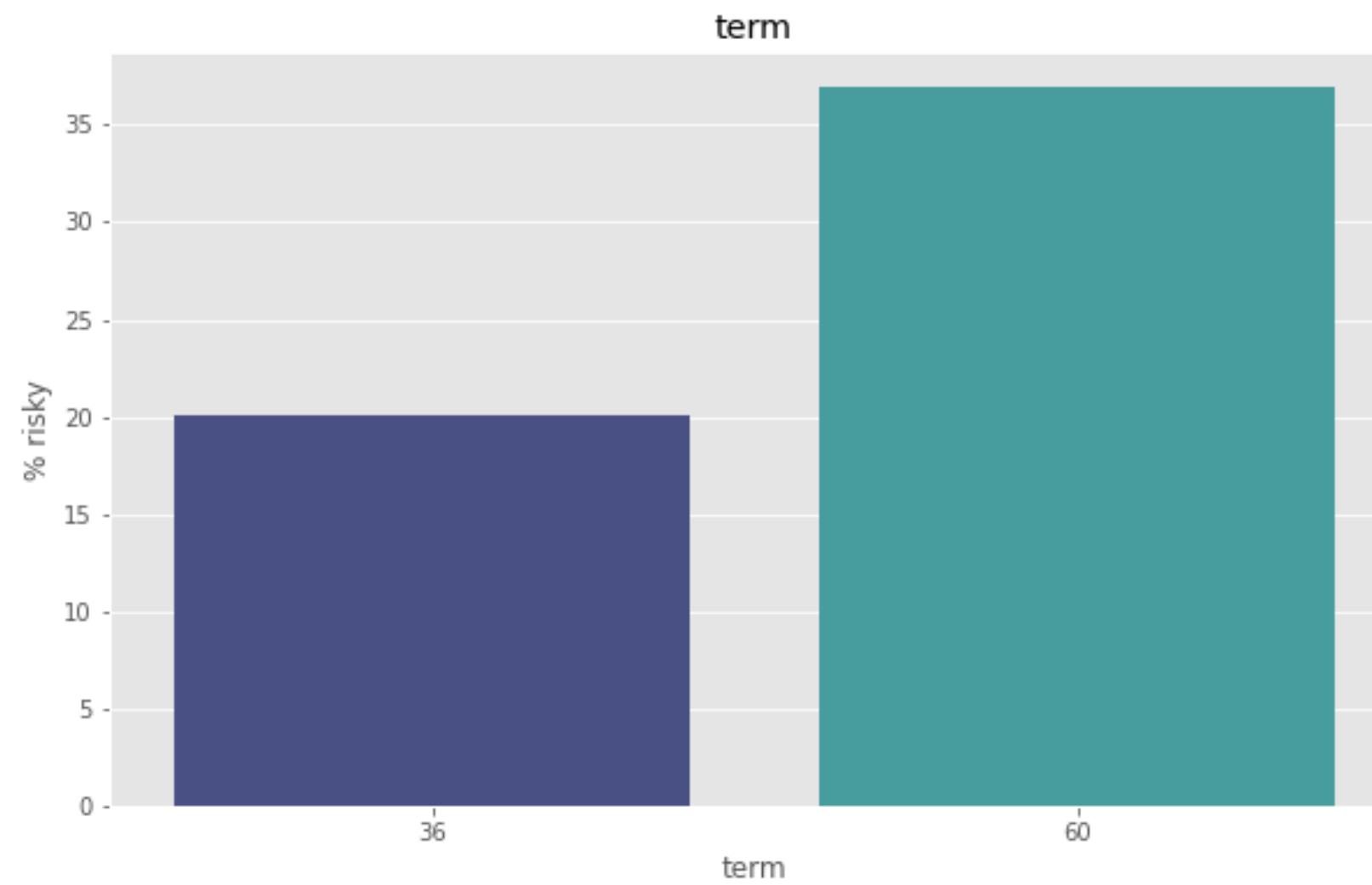


Dataset ini terdiri dari 74 kolom dan ~466rb data, untuk deskripsi fitur lebih lengkapnya dapat dilihat pada link berikut:

https://docs.google.com/spreadsheets/d/1iT1JNOBwU4I616_rnJpo0iny7blZvNBs/edit#gid=1666154857

DATA DESCRIPTION

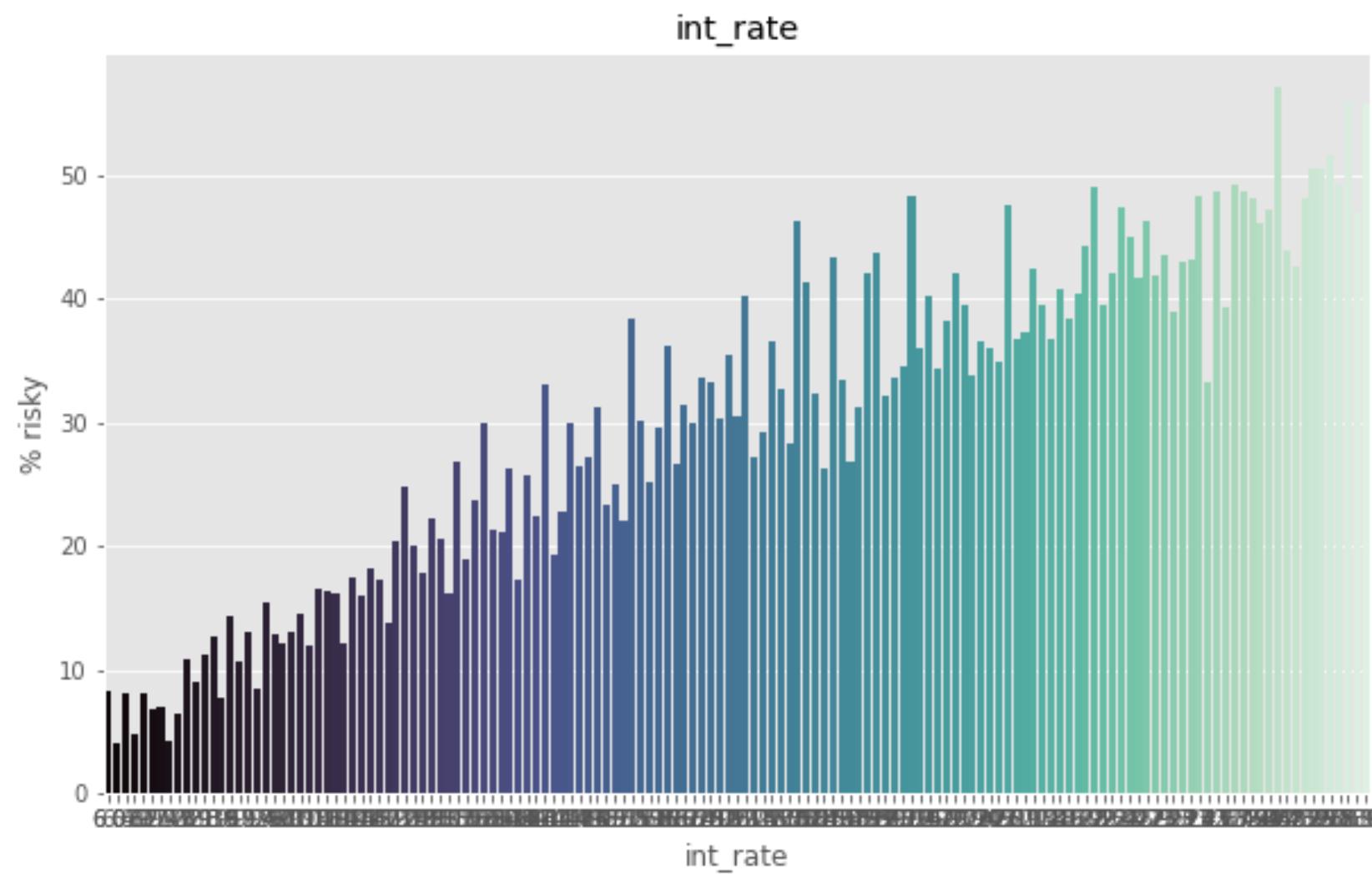
EXPLORATORY DATA ANALYSIS



Insight:

Risiko lebih tinggi terjadi pada peminjam dengan jangka waktu lebih dari 36 bulan.

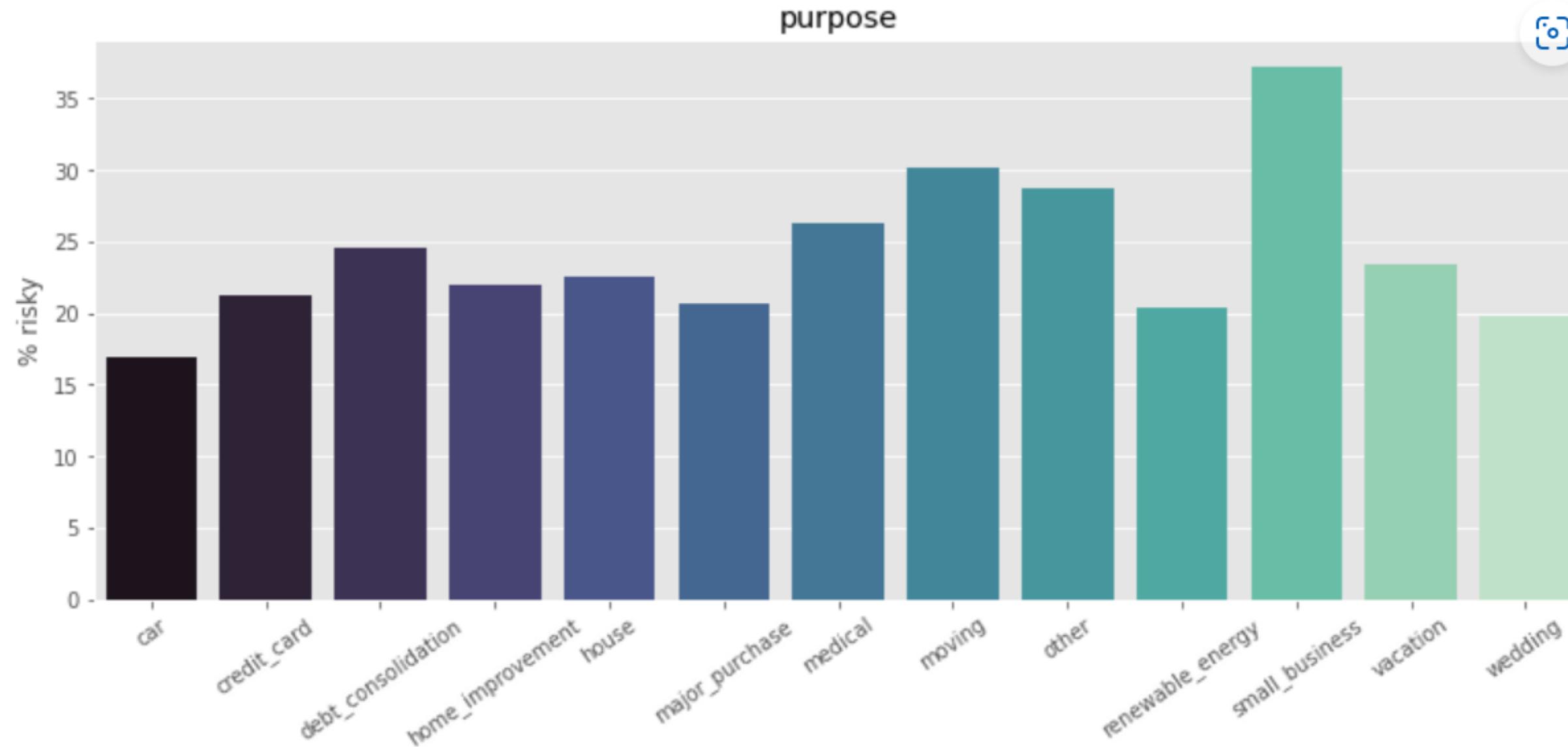
EXPLORATORY DATA ANALYSIS



Insight:

Semakin tinggi interest rate, maka akan semakin tinggi juga risiko kreditnya.

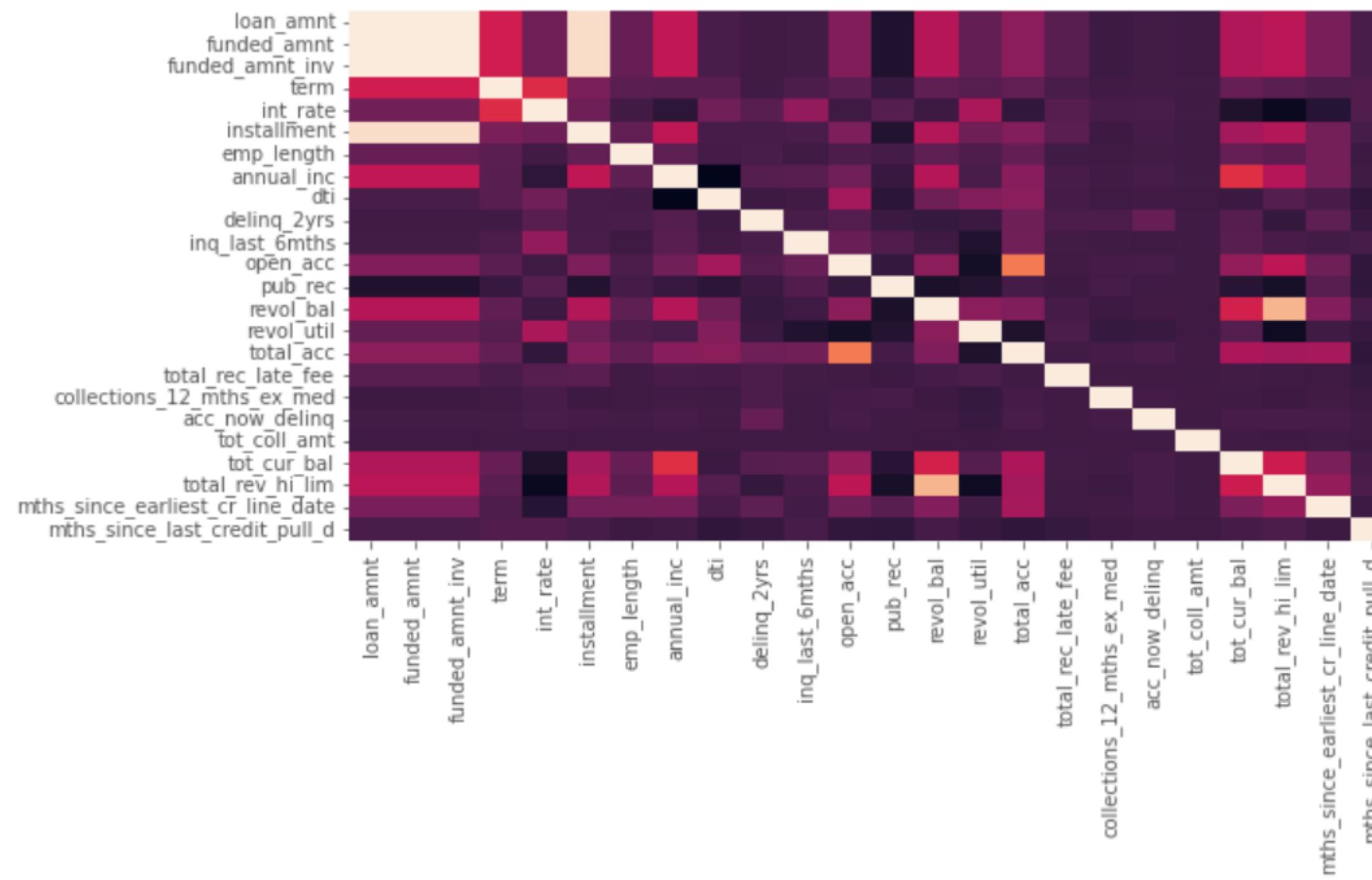
EXPLORATORY DATA ANALYSIS



Insight:

Peminjam berisiko biasanya memiliki tujuan meminjam untuk bisnis kecil, pindah, menutupi utang sebelumnya dan lainnya.

EXPLORATORY DATA ANALYSIS



Insight:

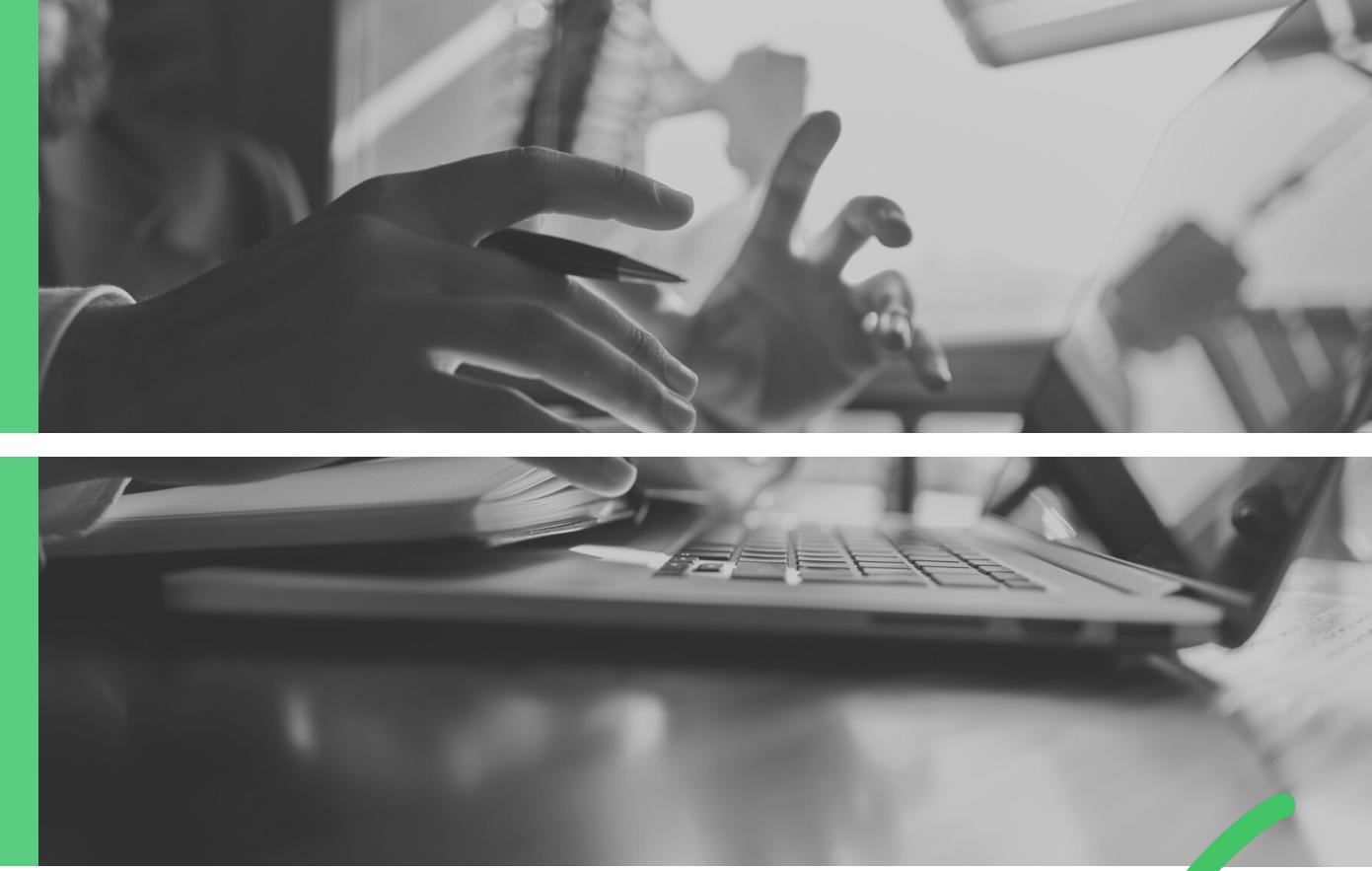
Kolom loan amount, funded amount dan funded amount inv memiliki korelasi yang tinggi, yang berarti data dari ketiga fitur tersebut mirip dan dapat diambil salah satu fitur saja.

PREPROCESSING

DATA CLEANING

- Menghapus kolom yang semua datanya null, data leakage, dan tidak dibutuhkan
- Mengisi null values dari kolom yang relevan dengan nilai yang sesuai dengan distribusi data
- Menambahkan fitur baru dari kolom yang sudah ada sebelumnya

#	Column	Non-Null Count	Dtype
0	loan_amnt	172290 non-null	int64
1	funded_amnt	172290 non-null	int64
2	funded_amnt_inv	172290 non-null	float64
3	term	172290 non-null	int64
4	int_rate	172290 non-null	float64
5	installment	172290 non-null	float64
6	emp_length	172290 non-null	int32
7	annual_inc	172290 non-null	float64
8	dti	172290 non-null	float64
9	delinq_2yrs	172290 non-null	float64
10	inq_last_6mths	172290 non-null	float64
11	open_acc	172290 non-null	float64
12	pub_rec	172290 non-null	float64
13	revol_bal	172290 non-null	int64
14	revol_util	172290 non-null	float64
15	total_acc	172290 non-null	float64
16	total_rec_late_fee	172290 non-null	float64
17	collections_12_mths_ex_med	172290 non-null	float64
18	acc_now_delinq	172290 non-null	float64
19	tot_coll_amt	172290 non-null	float64
20	tot_cur_bal	172290 non-null	float64
21	total_rev_hi_lim	172290 non-null	float64
22	mths_since_earliest_cr_line_date	172290 non-null	float64
23	mths_since_last_credit_pull_d	172290 non-null	float64



PREPROCESSING

ENCODING

- Membuat variabel dummy dengan menggunakan one-hot encoder pada kolom kategorikal

```
from sklearn.preprocessing import OneHotEncoder
cat_cols = ['grade', 'home_ownership', 'verification_status', 'purpose', 'initial_list_status']
onehot_cols = pd.get_dummies(df_cat[cat_cols])
```

```
onehot_cols.info()
```

```
...
```

```
onehot_cols.head()
```

```
...
```

```
onehot_cols['target'] = np.where(df_cat['good_risky_loan'].str.contains("good"), 1, 0)
```

```
...
```



PREPROCESSING

STANDARDIZATION

- Melakukan normalisasi dengan menggunakan standard scaler pada kolom numerikal

```
from sklearn.preprocessing import StandardScaler  
  
con_cols = [col for col in df_con.columns.tolist()]  
ss = StandardScaler()  
std_cols = pd.DataFrame(ss.fit_transform(df_con[con_cols]), columns=con_cols)
```

```
std_cols.info()
```

```
...  
...  
...  
...
```

```
std_cols.head()
```

```
...  
...
```



“



MODELING USING CLASSIFICATION ALGORITHMS

70%

DARI DATA UNTUK DILATIH
(TRAINING)

30%

DARI DATA UNTUK DIUJI
(TESTING)

LOGISTIC REGRESSION

```
# Logistic regression
lr= LogisticRegression(max_iter=600)
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
print('Classification_Report:')
print(classification_report(y_test, y_pred_lr))
```

	precision	recall	f1-score	support
risky loan	0.51	0.02	0.05	12452
good loan	0.76	0.99	0.86	39235
accuracy			0.76	51687
macro avg	0.64	0.51	0.45	51687
weighted avg	0.70	0.76	0.67	51687

	feature	importance
0	grade_A	0.445632
1	grade_B	0.263010
8	home_ownership_MORTGAGE	0.180100
13	verification_status_Not Verified	0.156075
29	initial_list_status_f	0.150682
30	initial_list_status_w	0.121207
18	purpose_debt_consolidation	0.098108
17	purpose_credit_card	0.078790
15	verification_status_Verified	0.075513
34	installment	0.069145
24	purpose_other	0.068059
11	home_ownership_OWN	0.055831
19	purpose_home_improvement	0.049719
12	home_ownership_RENT	0.049012
16	purpose_car	0.047659
32	term	0.046118

RANDOM FOREST CLASSIFIER

```
# random forest classifier
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
y_pred_rf = rf.predict(X_test)
print('Classification_Report:')
print(classification_report(y_test, y_pred_rf))

Classification_Report:
      precision    recall  f1-score   support

risky loan       0.48     0.07     0.12    12452
good loan        0.77     0.98     0.86    39235

accuracy          -         -     0.76    51687
macro avg       0.63     0.52     0.49    51687
weighted avg     0.70     0.76     0.68    51687
```

	feature	importance
37	dti	0.060159
49	tot_cur_bal	0.059906
43	revol_util	0.059089
42	revol_bal	0.058624
51	mths_since_earliest_cr_line_date	0.057895
50	total_rev_hi_lim	0.057371
34	installment	0.057349
36	annual_inc	0.054320
33	int_rate	0.052294
44	total_acc	0.048863
31	loan_amnt	0.047877
40	open_acc	0.041392
52	mths_since_last_credit_pull_d	0.033548
35	emp_length	0.030637
0	grade_A	0.029227

SUMMARY

Dari kedua model klasifikasi yang telah dibuat menggunakan Logistic Regression dan Random Forest Classifier, diperoleh akurasi yang sama yaitu **76%** dan tergolong baik. Jadi, model ini dapat digunakan untuk memprediksi data baru dalam rangka pengambilan keputusan apakah pengajuan peminjam tersebut dapat diterima atau ditolak berdasarkan tingkat risikonya.

Kesalahan prediksi dari kedua model tersebut dipengaruhi oleh feature importance. Untuk model Logistic Regression, fitur yang paling penting adalah fitur kelas, kepemilikan rumah dan status verifikasi. Sedangkan, untuk model Random Forest Classifier, fitur yang paling penting adalah rasio antara total pembayaran utang bulanan peminjam dengan pendapatan bulanan peminjam, dan total saldo dari semua akun peminjam tersebut.



THANK YOU



<https://www.linkedin.com/in/yemima-sipayung/>



<https://github.com/Yemimaaa>