

FINAL PROJECT

CREDIT RISK LOAN

A Machine Learning Model for Credit Risk classification

Project By
Muhammad Ravil

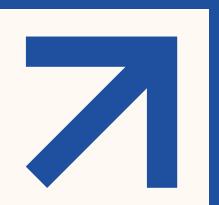


Project Based Internship - id/x partners

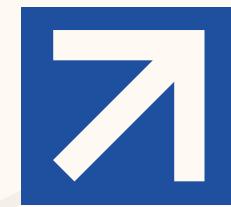


MUHAMMAD RAVIL

LULUSAN S1 TEKNIK INFORMATIKA DARI UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU DENGAN MINAT DALAM PENGEMBANGAN APLIKASI WEB, PEMECAHAN MASALAH TEKNIS (IT TROUBLESHOOTING), SERTA PENERAPAN MACHINE LEARNING UNTUK ANALISIS DAN PENGOLAHAN DATA SECARA EFEKTIF.



id/x partners



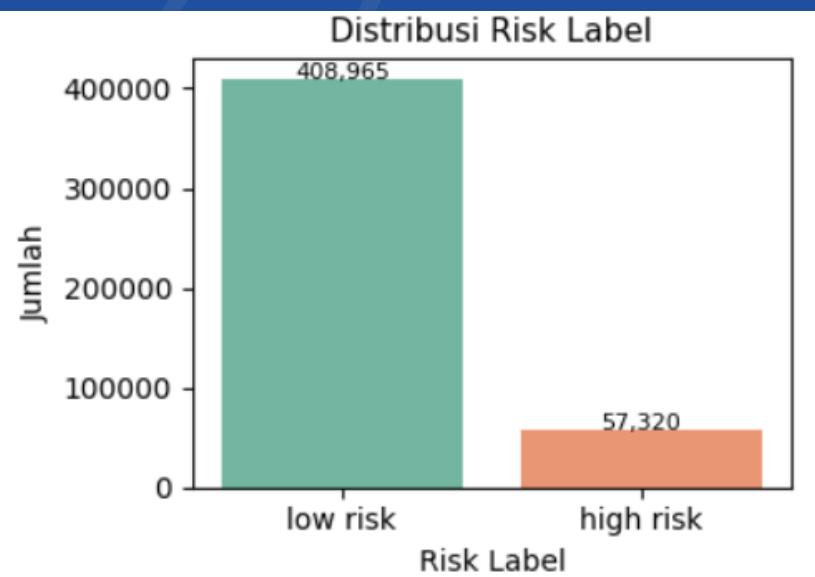
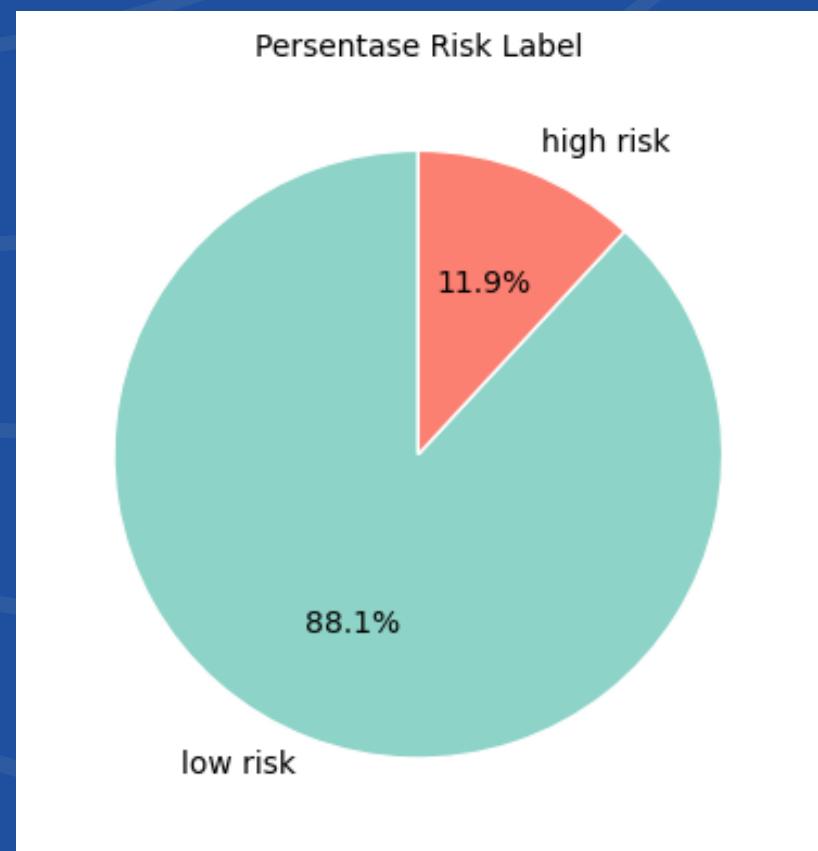
BUSSINES UNDERSTANDING



PROBLEM STATEMENT



BANK MENGHADAPI TANTANGAN BESAR DALAM MENILAI RISIKO KREDIT, TERCERMIN DARI TINGGINYA RASIO KREDIT MACET YANG MENCAPAI 12%. HAL INI BERDAMPAK LANGSUNG PADA STABILITAS KEUANGAN DAN EFISIENSI OPERASIONAL. SISTEM EVALUASI KREDIT YANG ADA SAATINI BELUM MAMPU MEMPREDIKSI RISIKO NASABAH SECARA AKURAT. DARI TOTAL 466.285 PEMINJAM, DISTRIBUSI STATUS PINJAMAN SANGAT BERAGAM — MULAI DARI PINJAMAN LANCAR, TELAH LUNAS, HINGGA DEFAULT BERISIKO TINGGI. VISUALISASI DATA BERIKUT MEMBERIKAN GAMBARAN MENYELURUH MENGENAI KONDISI INI, SEKALIGUS MENEGASKAN PERLUNYA PENERAPAN SOLUSI BERBASIS MACHINE LEARNING UNTUK MENINGKATKAN KETEPATAN PREDIKSI RISIKO KREDIT.



OBJECTIVE

- MEMPERKUAT KEBIJAKAN KREDIT DENGAN ANALISIS DATA YANG LEBIH AKURAT UNTUK MENINGKATKAN VALIDITAS KEPUTUSAN PEMINJAMAN.
- MENERAPKAN MODEL MACHINE LEARNING UNTUK MENILAI RISIKO PINJAMAN DAN MENURUNKAN PINJAMAN BERISIKO TINGGI.

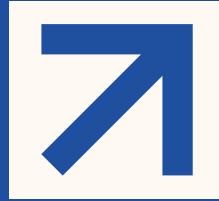
GOAL

MEMBANGUN MODEL PREDIKSI UNTUK MENGURANGI TINGKAT BAD LOAN SERTA MEMPERKUAT VALIDITAS DAN AKURASI KEPUTUSAN PEMINJAMAN.



DATA UNDERSTANDING (EDA)





DATASET OVERVIEW

1. DATA MEMILIKI 466.285 BARIS DAN 74 KOLOM
2. DATASET TERDIRI DARI 3 JENIS TIPE DATA: INT64, FLOAT64, DAN OBJECT.
3. FITUR ISSUE_D, LAST_PYMNT_D, NEXT_PYMNT_D, LAST_CREDIT_PULL_D, DAN EARLIEST_CR_LINE HARUS DIUBAH MENJADI TIPE DATA DATETIME.
4. ADA 17 KOLOM YANG TIDAK MEMILIKI NILAI SAMA SEKALI
5. ADA 40 KOLOM YANG MEMILIKI NILAI NULL
6. IMBALANCE CLASS YG SANGAT TINGGI ANTARA LABEL LOW RISK DAN HIGH RISK
7. TIDAK TERDAPAT DUPLIKASI BARIS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 75 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Unnamed: 0       466285 non-null   int64  
 1   id               466285 non-null   int64  
 2   member_id        466285 non-null   int64  
 3   loan_amnt        466285 non-null   int64  
 4   funded_amnt      466285 non-null   int64  
 5   funded_amnt_inv  466285 non-null   float64
 6   term              466285 non-null   object  
 7   int_rate          466285 non-null   float64
 8   installment       466285 non-null   float64
 9   grade             466285 non-null   object  
 10  sub_grade         466285 non-null   object  
 11  emp_title         438697 non-null   object  
 12  emp_length        445277 non-null   object  
 13  home_ownership    466285 non-null   object  
 14  annual_inc        466281 non-null   float64
 15  verification_status 466285 non-null   object  
 16  issue_d           466285 non-null   object  
 17  loan_status        466285 non-null   object  
 18  pymnt_plan         466285 non-null   object  
 19  url               466285 non-null   object  
 20  desc               125981 non-null   object  
 21  purpose            466285 non-null   object  
 22  title              466264 non-null   object  
 23  zip_code           466285 non-null   object  
 24  addr_state         466285 non-null   object  
 25  dti                466285 non-null   float64
 26  delinq_2yrs        466256 non-null   float64
 27  earliest_cr_line   466256 non-null   object  
 28  inq_last_6mths     466256 non-null   float64
 29  mths_since_last_delinq 215934 non-null   float64
 30  mths_since_last_record 62638 non-null   float64
 31  open_acc           466256 non-null   float64
 32  pub_rec             466256 non-null   float64
 33  revol_bal           466285 non-null   int64  
 34  revol_util          465945 non-null   float64
 35  total_acc            466256 non-null   float64
 36  initial_list_status 466285 non-null   object  
 37  out_prncp           466285 non-null   float64
 38  out_prncp_inv       466285 non-null   float64
 39  total_pymnt         466285 non-null   float64
 40  total_pymnt_inv     466285 non-null   float64
 41  total_rec_prncp     466285 non-null   float64
 42  total_rec_int        466285 non-null   float64
 43  total_rec_late_fee   466285 non-null   float64
 44  recoveries           466285 non-null   float64
 45  collection_recovery_fee 466285 non-null   float64
 46  last_pymnt_d         465909 non-null   object  
 47  last_pymnt_amnt      466285 non-null   float64
 48  next_pymnt_d         239871 non-null   object  
 49  last_credit_pull_d    466243 non-null   object  
 50  collections_12_mths_ex_med 466148 non-null   float64
 51  mths_since_last_major_derog 98974 non-null   float64
 52  policy_code           466285 non-null   int64  
 53  application_type      466285 non-null   object 
```

```
53  application_type      466285 non-null   object  
54  annual_inc_joint      0 non-null      float64
55  dti_joint              0 non-null      float64
56  verification_status_joint 0 non-null      float64
57  acc_now_delinq         466256 non-null   float64
58  tot_coll_amt           396889 non-null   float64
59  tot_cur_bal             396889 non-null   float64
60  open_acc_6m             0 non-null      float64
61  open_il_6m              0 non-null      float64
62  open_il_12m             0 non-null      float64
63  open_il_24m             0 non-null      float64
64  mths_since_rcnt_il     0 non-null      float64
65  total_bal_il            0 non-null      float64
66  il_util                 0 non-null      float64
67  open_rv_12m              0 non-null      float64
68  open_rv_24m              0 non-null      float64
69  max_bal_bc              0 non-null      float64
70  all_util                 0 non-null      float64
71  total_rev_hi_lim       396889 non-null   float64
72  inq_fi                  0 non-null      float64
73  total_cu_tl              0 non-null      float64
74  inq_last_12m             0 non-null      float64
dtypes: float64(46), int64(7), object(22)
memory usage: 266.8+ MB
```



TARGET VARIABEL

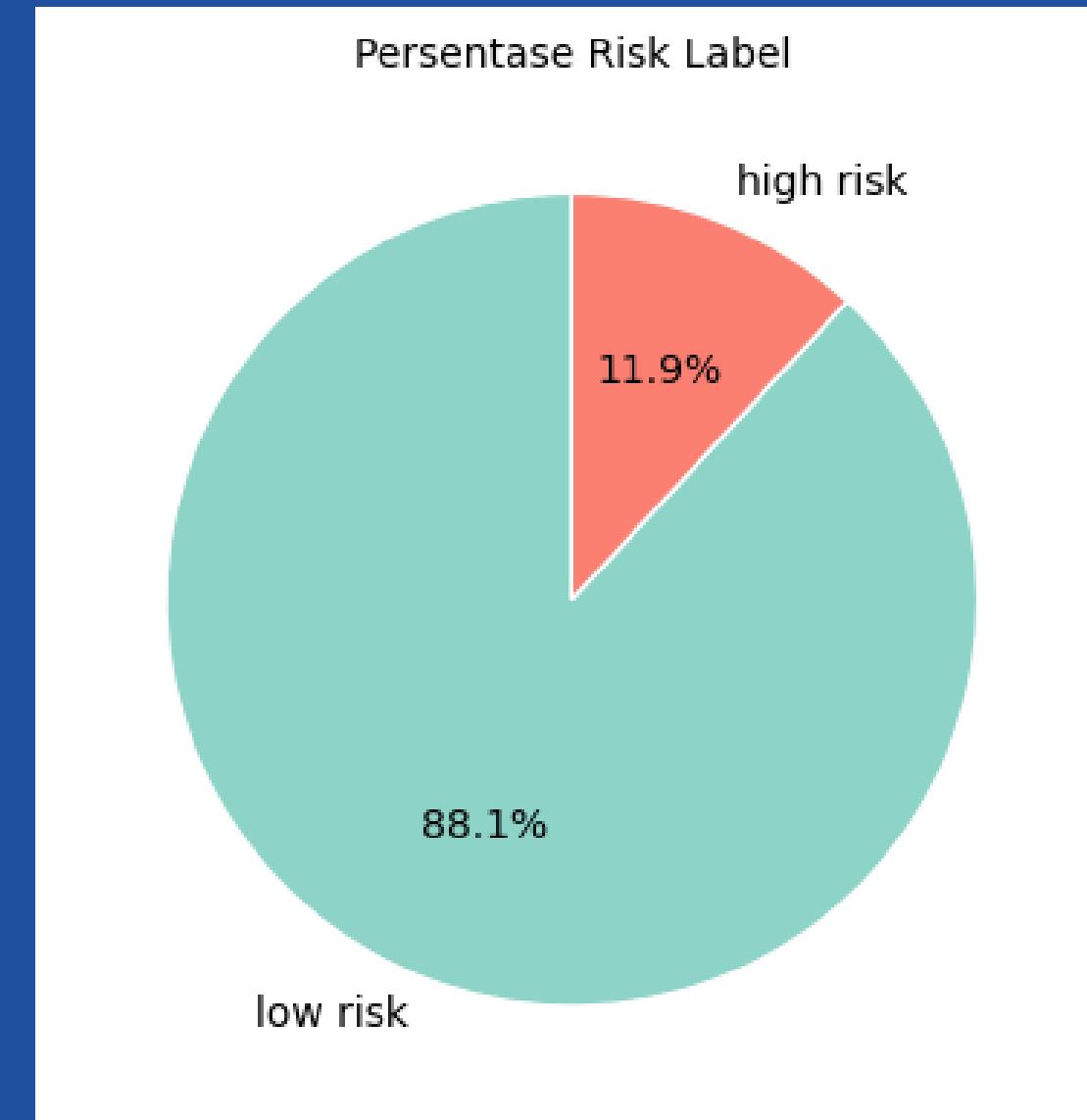
VARIABEL TARGET AKAN DIKELOMPOKKAN DARI KOLOM LOAN_STATUS, BERDASARKAN KONDISI TERTENTU YANG AKAN DITENTUKAN.

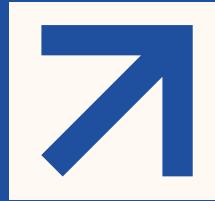
LOW-RISK LOAN STATUS:

- FULLY PAID
- CURRENT
- DOES NOT MEET THE CREDIT POLICY. STATUS:FULLY PAID

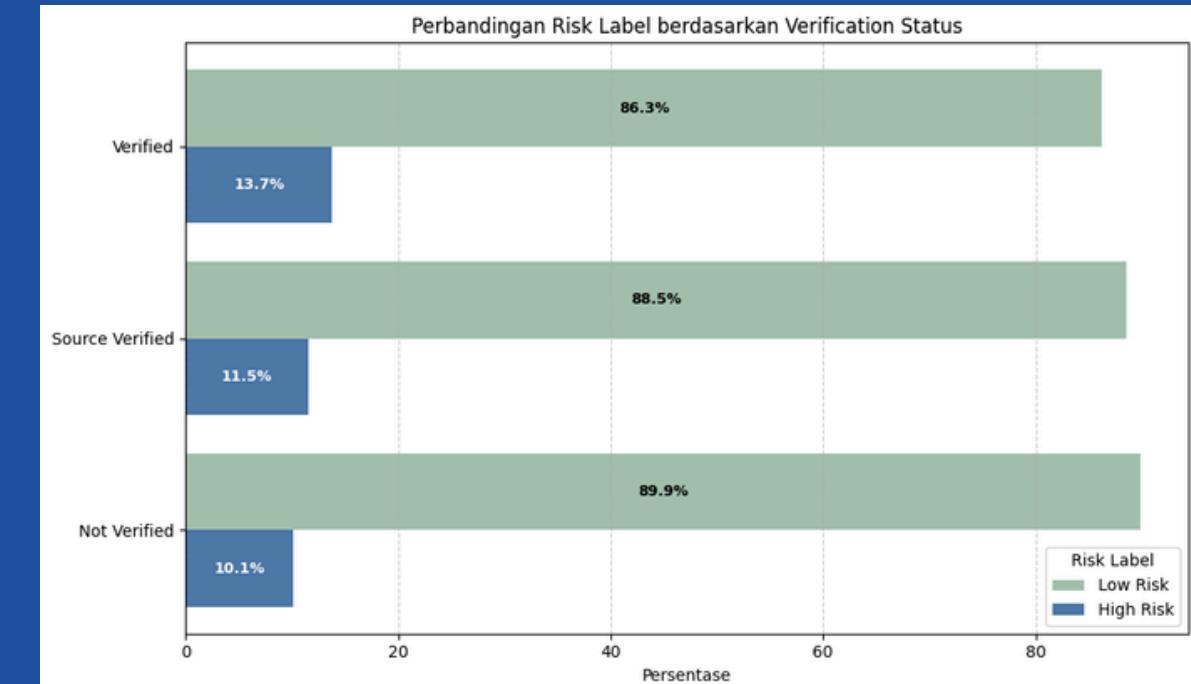
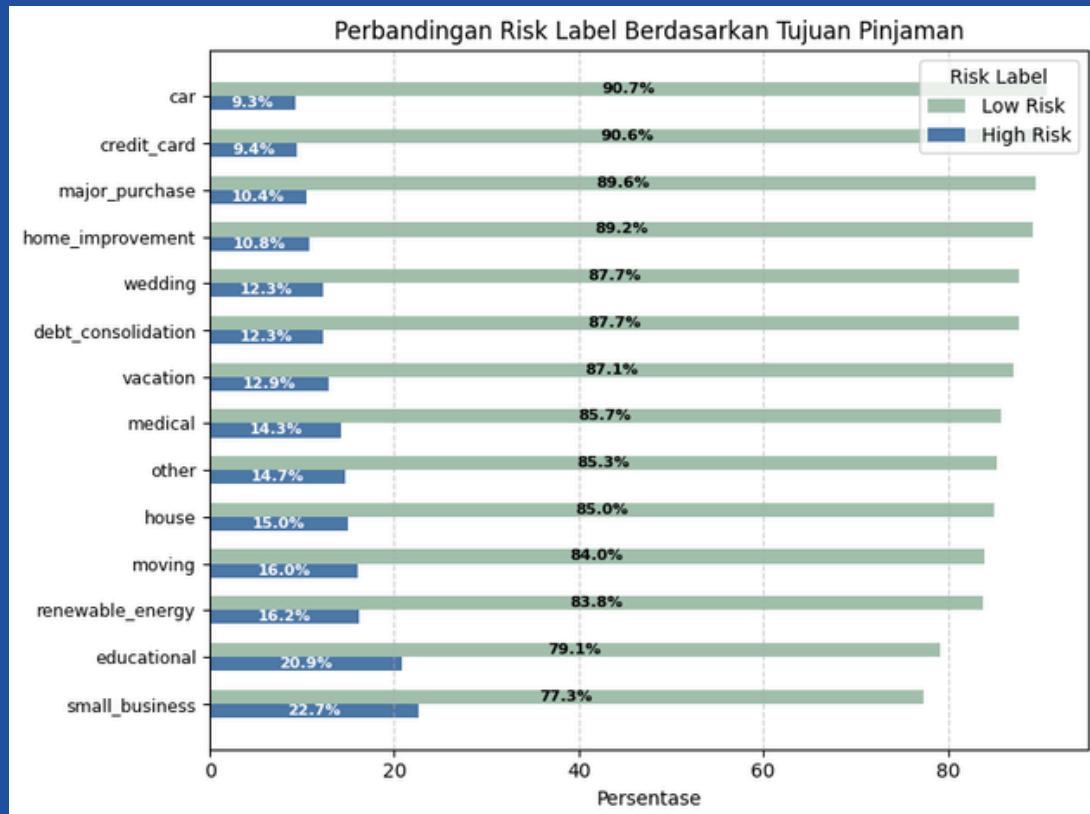
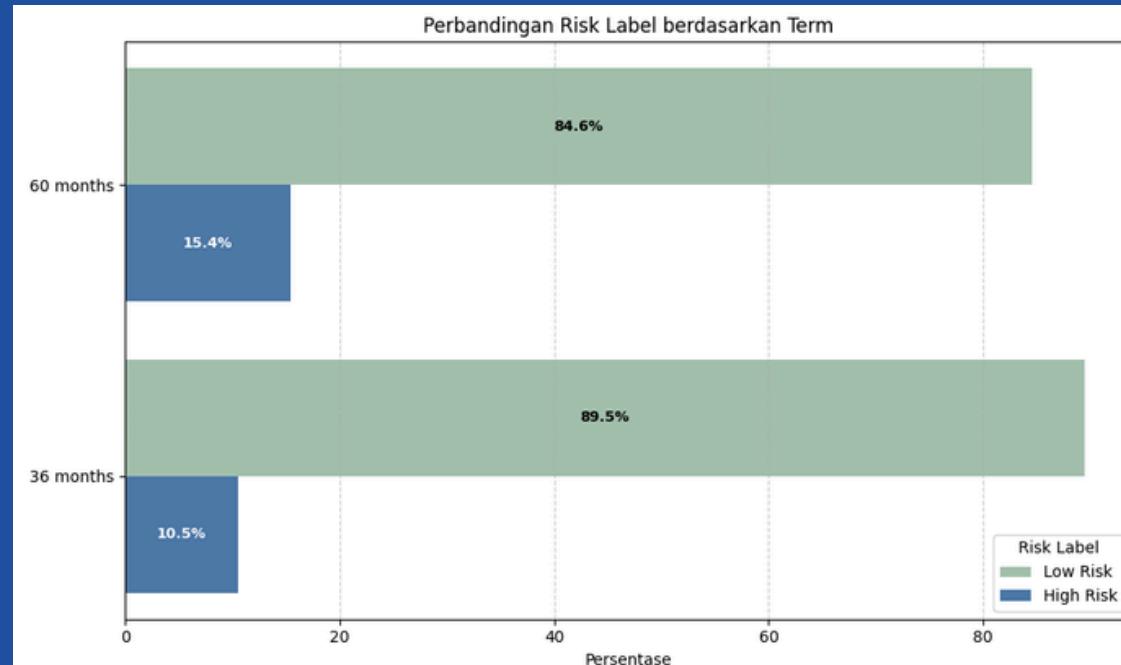
HIGH-RISK LOAN STATUS:

- LATE (16 - 30 DAYS)
- LATE (31 - 120 DAYS)
- DEFAULT
- CHARGED OFF
- IN GRACE PERIOD
- DOES NOT MEET THE CREDIT POLICY. STATUS:CHARGED OF





VARIATE ANALYSIS



TERM

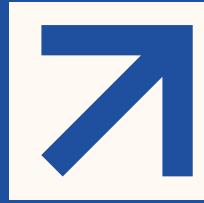
SEMAKIN PANJANG JANGKA WAKTU PINJAMAN, SEMAKIN BESAR PROPORSI RISIKO TINGGI. PADA TENOR 36 BULAN, 10,5% TERMASUK RISIKO TINGGI, SEDANGKAN PADA 60 BULAN MENINGKAT MENJADI 15,4%. SEBALIKNYA, PROPORSI RISIKO RENDAH MENURUN DARI 89,5% MENJADI 84,6%.

PURPOSE

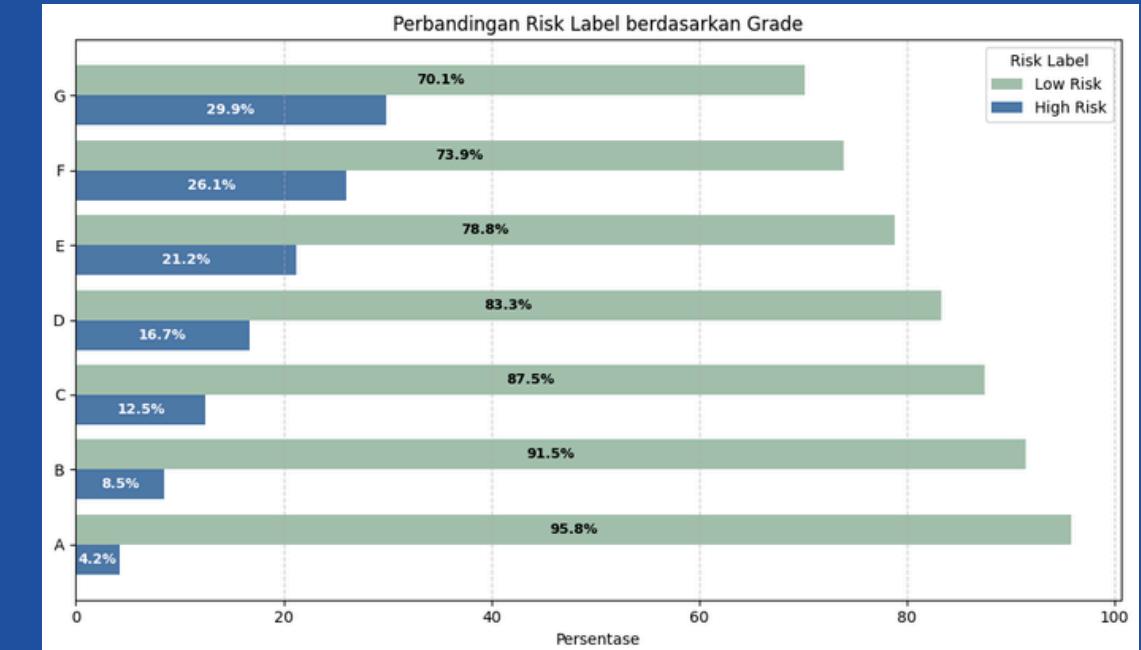
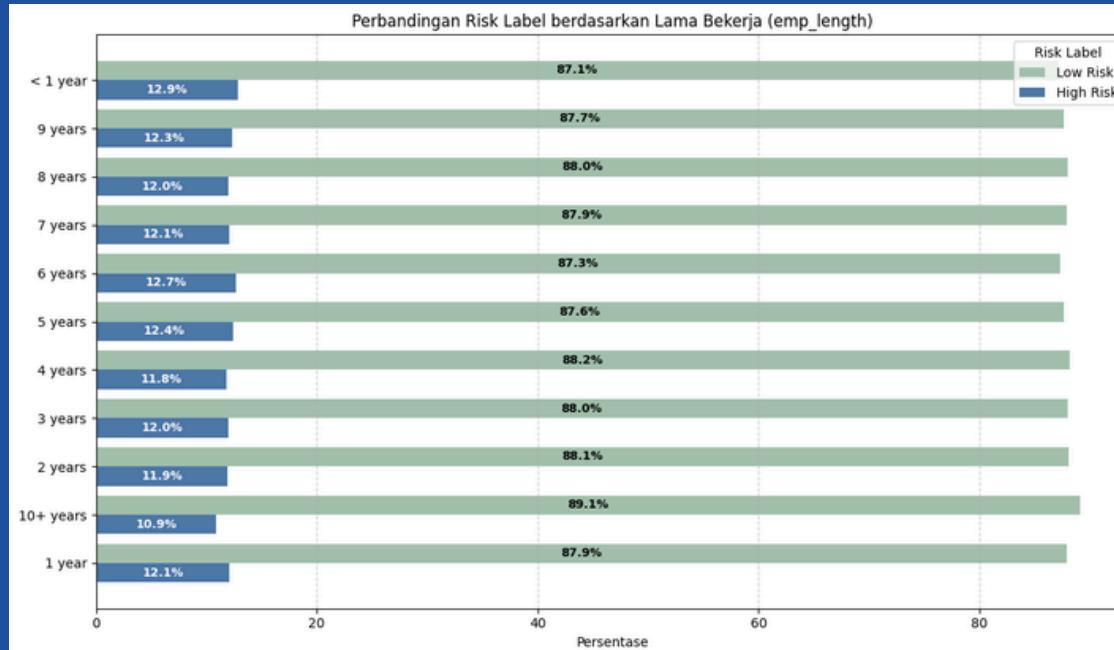
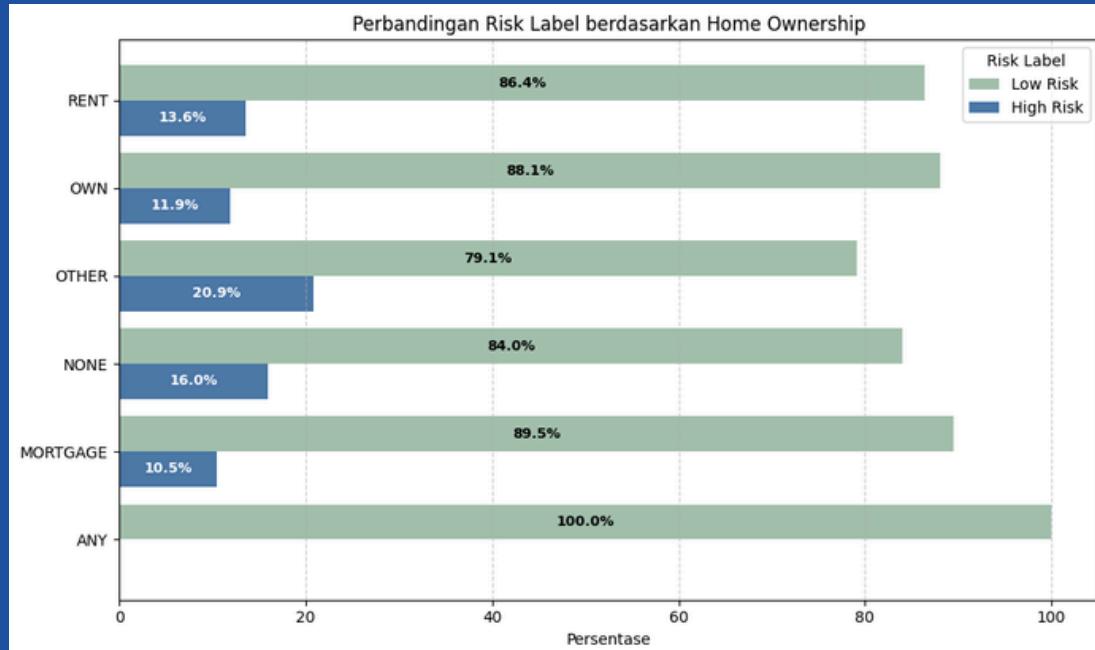
PINJAMAN DENGAN TUJUAN SEPERTI CAR DAN CREDIT_CARD TERGOLONG PALING AMAN, DENGAN RISIKO RENDAH DI ATAS 90%. SEBALIKNYA, PINJAMAN UNTUK SMALL_BUSINESS DAN EDUCATIONAL MENUNJUKKAN TINGKAT RISIKO TINGGI YANG SIGNIFIKAN, MASING-MASING MENCAPAI 22,7% DAN 20,9%. SECARA KESELURUHAN, PINJAMAN KONSUMTIF CENDERUNG LEBIH RENDAH RISIKONYA DIBANDINGKAN DENGAN PINJAMAN PRODUKTIF SEPERTI USAHA ATAU PENDIDIKAN.

VERIFICATION STATUS

KLIEN DENGAN STATUS VERIFIKASI "NOT VERIFIED" MEMILIKI TINGKAT GAGAL BAYAR TERTINGGI SEBESAR 20,19%, SEMENTARA "SOURCE VERIFIED" MENUNJUKKAN TINGKAT TERENDAH SEBESAR 16%. HAL INI MENGINDIKASIKAN BAHWA PROSES VERIFIKASI BERPERAN DALAM MENINGKATKAN KUALITAS DAN KELAYAKAN KREDIT.



VARIATE ANALYSIS



HOME OWNER

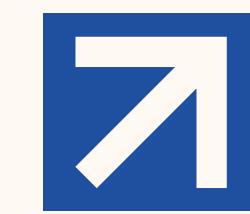
SECARA KESELURUHAN, STATUS KEPEMILIKAN RUMAH MEMENGARUHI TINGKAT RISIKO, DENGAN "MORTGAGE" SEBAGAI YANG PALING AMAN DAN "NONE" SEBAGAI YANG PALING BERISIKO DI ANTARA KATEGORI YANG DITAMPILKAN.

EMP LENGTH

SEMAKIN PANJANG DURASI MASA KERJA, SEMAKIN KECIL KEMUNGKINAN KLIEN GAGAL MEMBAYAR PINJAMAN. KLIEN DENGAN PENGALAMAN KERJA KURANG DARI SATU TAHUN MEMILIKI TINGKAT GAGAL BAYAR TERTINGGI SEBESAR 21,42%, YANG MENGINDIKASIKAN BAHWA KESTABILAN PEKERJAAN BERPERAN PENTING DALAM MENJAGA KELAYAKAN KREDIT.

GRADE

GRAFIK INI MENUNJUKKAN BAHWA GRADE "A" MEMILIKI TINGKAT LOW RISK TERTINGGI SEBESAR 95,8%, YANG BERARTI PINJAMAN DENGAN GRADE INI PALING AMAN. SEBALIKNYA, GRADE "G" MEMILIKI TINGKAT HIGH RISK TERTINGGI SEBESAR 29,9%, MENANDAKAN BAHWA PINJAMAN DENGAN GRADE INI PALING BERISIKO.



DATA PREPROCESSING



DATA PREPROCESSING

id/x partners



DATA CLEANING

NILAI YANG HILANG DITANGANI DENGAN MENGHAPUS DATA ATAU MENGIMPUTASI MENGGUNAKAN MODE ATAU MEDIAN.



LABEL TARGET

LOW-RISK (1): FULLY PAID, CURRENT, IN GRACE PERIOD. HIGH-RISK (0): DEFAULT, CHARGED OFF, LATE (16–120 DAYS), AND DOES NOT MEET CREDIT POLICY.



FEATURE ENCODING

MENERAPKAN MANUAL, BINARY, ORDINAL, DAN TARGET ENCODING UNTUK FITUR KATEGORIKAL.



FEATURE ENGINEERING

KOLOM TANGGAL DIUBAH MENJADI TAHUN (ISSUE_D_YEAR) DAN (CREDIT_AGE). KATEGORI HOME_OWNERSHIP DISEDERHANAKAN, MENGGABUNGKAN 'ANY' DAN 'NONE' MENJADI 'OTHER', SERTA KATEGORI PURPOSE DIKELOMPOKKAN KE DALAM 'PRIVATE_USE'.



DATA PREPROCESSING

id/x partners



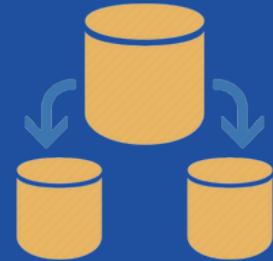
FEATURE SELECTION

MENGURANGI TOTAL DARI 74 MENJADI HANYA 20 KOLOM DENGAN MENGHAPUS FITUR YANG MEMILIKI KORELASI TINGGI >0,80.



FEATURE SCALING

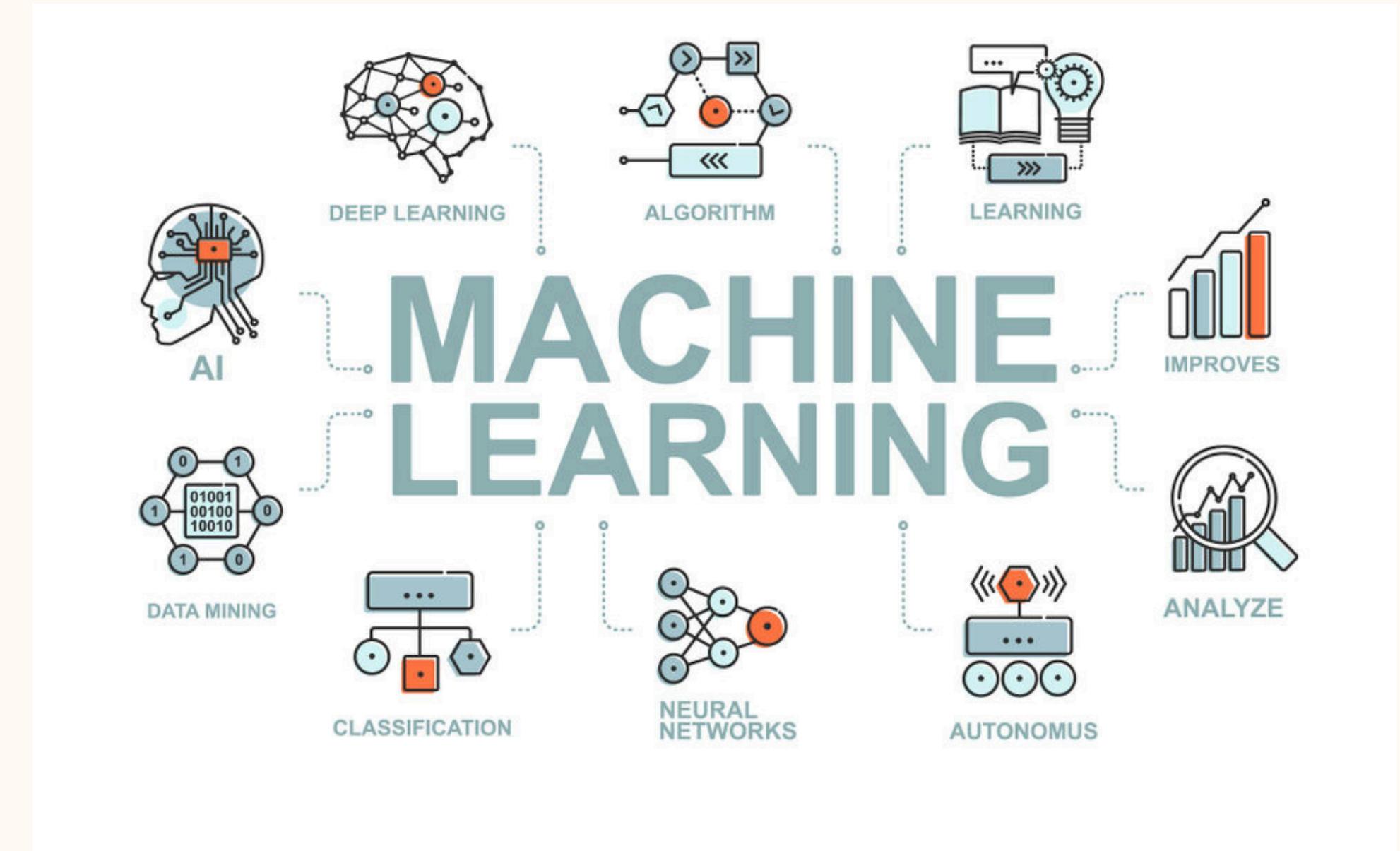
MENGGUNAKAN MINMAXSCALER UNTUK MENSKALAKAN FITUR NUMERIK



DATA SPLIT & DATA BALANCING

MEMBAGI DATA DENGAN RASIO 80:20, MENGGUNAKAN 5-FOLD CV DAN SMOTE UNTUK MENYEIMBANGKAN KELAS MENJADI 50:50.

MODELLING



DARI BEBERAPA MODEL YANG DIUJI, RANDOM FOREST DENGAN SMOTE MENUNJUKKAN PERFORMA TERBAIK DENGAN F1_TEST SEBESAR 0.963 (ATAU SEKITAR 96.3%, MENDEKATI 98%), MENGUNGULI LOGISTIC REGRESSION DAN DECISION TREE. LOGISTIC REGRESSION MEMILIKI PERFORMA BAIK NAMUN SEDIKIT LEBIH RENDAH DENGAN AKURASI 0.950 DAN CROSSVAL_AUC 0.974, SEDANGKAN DECISION TREE MENUNJUKKAN TANDA-TANDA OVERTFITTING DENGAN PERBEDAAN YANG SIGNIFIKAN ANTARA AUC_TRAIN 1.000 DAN AUC_TEST 0.927 SEBELUM SMOTE. OLEH KARENA ITU, RANDOM FOREST DENGAN SMOTE DIIMPLEMENTASIKAN SEBAGAI MODEL TERBAIK UNTUK ANALISIS PREDIKSI INI.

TANPA SMOTE

Model	Accuracy	Model Comparison (Train, Test, and CrossVal)								
		AUC_train	AUC_test	Recall_train	Recall_test	Precision_train	Precision_test	F1_train	F1_test	CrossVal_AUC
0 Random Forest	0.934	1.000	0.848	1.000	0.999		1.000	0.931	1.000	0.964 0.845
1 Logistic Regression	0.914	0.814	0.815	0.998	0.998		0.913	0.912	0.954	0.953 0.812
2 Decision Tree	0.883	1.000	0.737	1.000	0.929		1.000	0.938	1.000	0.933 0.733

DENGAN SMOTE

Model	Accuracy	Model Comparison (Train, Test, and CrossVal)								
		AUC_train	AUC_test	Recall_train	Recall_test	Precision_train	Precision_test	F1_train	F1_test	CrossVal_AUC
0 Random Forest	0.961	1.000	0.983	1.000	0.995		1.000	0.932	1.000	0.963 0.982
1 Logistic Regression	0.950	0.974	0.974	0.997	0.997		0.912	0.911	0.953	0.952 0.974
2 Decision Tree	0.927	1.000	0.927	1.000	0.921		1.000	0.932	1.000	0.927 0.925

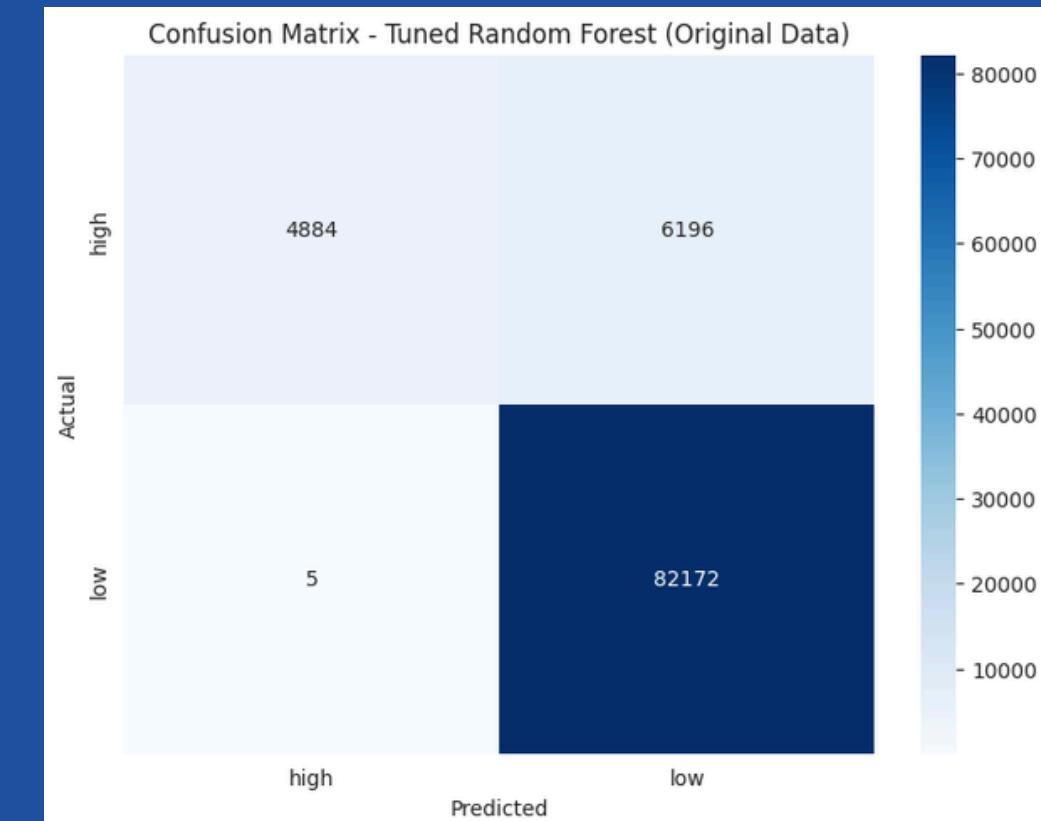
TUNING MODEL

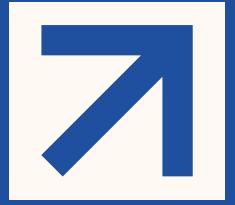
SETELAH DIPEROLEH BAHWA ALGORITMA TERBAIK ADALAH RANDOM FOREST, SELANJUTNYA AKAN DILAKUKAN TUNING HYPERPARAMETER PADA MODEL INI UNTUK MENGETAHUI PARAMETER APA SAJA YANG DAPAT MENINGKATKAN AKURASI MODEL.

PARAMETER YANG DI UJI:

- N_ESTIMATORS: [100-200] — JUMLAH POHON; LEBIH BANYAK POHON MENINGKATKAN STABILITAS, TAPI MEMPERLAMA KOMPUTASI.
- MAX_DEPTH: [10, 20, 30, NONE] — BATAS KEDALAMAN POHON UNTUK MENGONTROL KOMPLEKSITAS DAN OVERFITTING.
- MAX_FEATURES: ['SQRT', 'LOG2', NONE] — JUMLAH FITUR SAAT SPLIT; MEMBANTU GENERALISASI MODEL.
- MIN_SAMPLES_SPLIT: [2-20] — MINIMUM SAMPEL UNTUK MEMBAGI NODE; MENCEGAH OVERFITTING.
- MIN_SAMPLES_LEAF: [1-10] — MINIMUM SAMPEL DI SETIAP DAUN; MENJAGA MODEL TETAP SEDERHANA.
- MAX_LEAF_NODES: [NONE, 50, 100] — BATAS JUMLAH DAUN UNTUK MENGENDALIKAN KOMPLEKSITAS DAN WAKTU PELATIHAN.

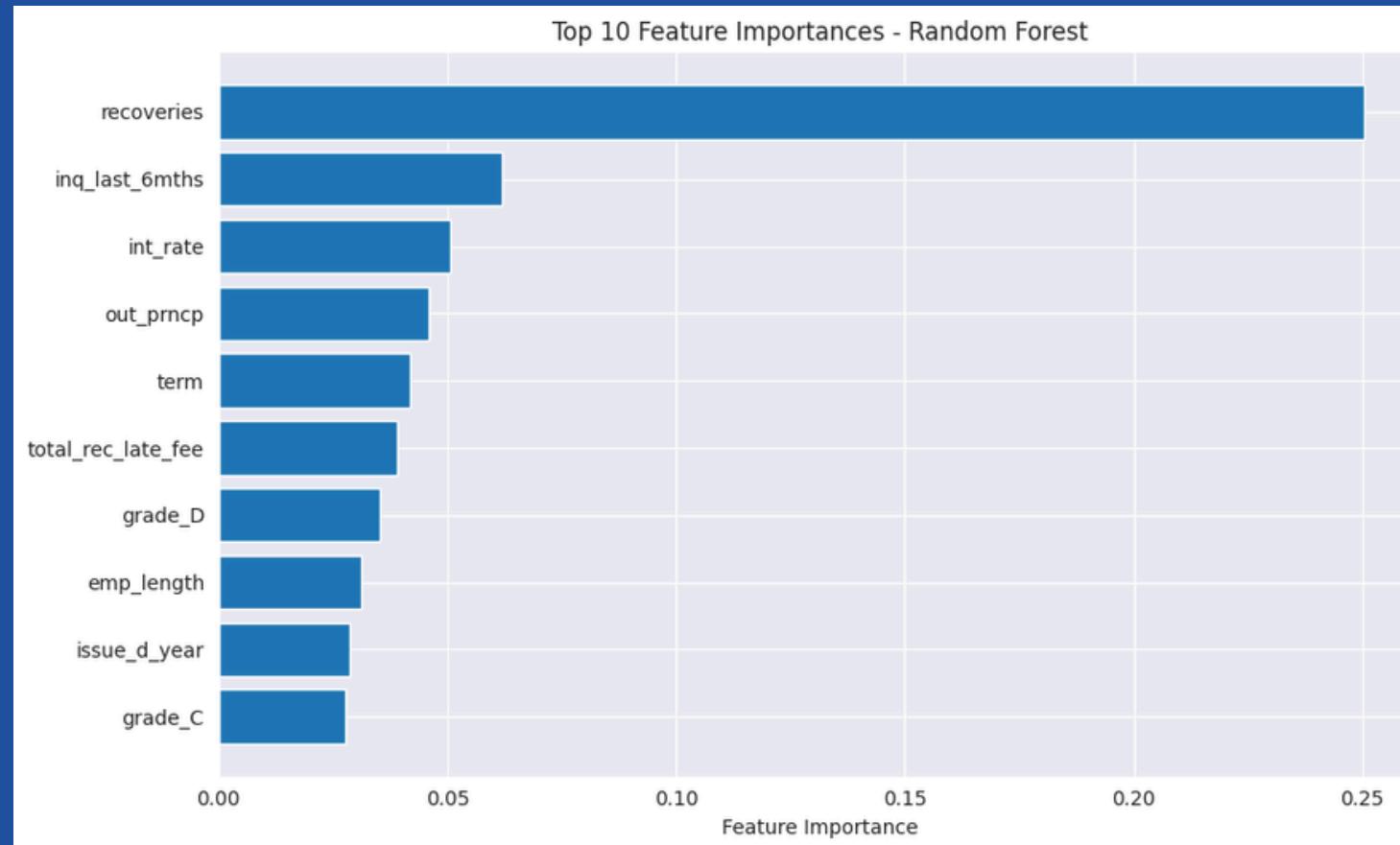
```
== Evaluation Metrics ==
Accuracy : 0.9572
AUC Train : 0.9938
AUC Test : 0.9807
Recall Train : 0.9994
Recall Test : 0.9961
Precision Train : 0.9319
Precision Test : 0.9244
F1 Score Train : 0.9645
F1 Score Test : 0.9589
```





FEATURE IMPORTANCE

id/x partners



REKOMENDASI BISNIS

1. RECOVERIES : BANGUN TIM KHUSUS UNTUK MENANGANI AKUN BERMASALAH, TAWARKAN NEGOSIASI PEMBAYARAN BERTAHAP, DAN GUNAKAN TEKNOLOGI AI UNTUK MEMPREDIKSI PELANGGAN YANG BERPOTENSI PULIH.
2. INQ_LAST_6MTHS : BATASI PERSETUJUAN UNTUK PELAMAR DENGAN LEBIH DARI 2-3 PERTANYAAN DALAM 6 BULAN, ATAU BERI EDUKASI TENTANG DAMPAK PERTANYAAN KREDIT BERLEBIH UNTUK MENGURANGI RISIKO.
3. INTEREST RATE : TAWARKAN SUKU BUNGA KOMPETITIF
4. OUT_PRNCP : BERIKAN INSENTIF SEPERTI DISKON UNTUK PELUNASAN DINI.
5. TERM : PRIORITASKAN TENOR YANG OPTIMAL DENGAN TINGKAT KEBERHASILAN PENYELESAIAN YANG TINGGI.
6. GRADE : KENCANGKAN PERSYARATAN UNTUK GRADE
7. TOTAL_REC_LATE_FEE : TERAPKAN SISTEM PERINGATAN UNTUK PELANGGAN YANG MENDEKATI KETERLAMBATAN, SERTA TAWARKAN PROGRAM PENGHAPUSAN BIAYA JIKA MEMBAYAR TEPAT WAKTU.
8. EMP_LENGTH : BERI PRIORITAS ATAU INSENTIF KEPADA PELANGGAN DENGAN DENGAN MASA KERJA YANG LAMA
9. ISSUE_D_YEAR : ANALISIS "ISSUE_D" HISTORIS UNTUK MENGIDENTIFIKASI POLA RISIKO

THANK YOU



LINK FINAL TASK

LINK FOLDER (VIDEO,CODING):

[HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1OKXTJ6V3MHK7HWWIBHTA1HVKSCY45WZ_?USP=DRIVE_LINK](https://drive.google.com/drive/folders/1OKXTJ6V3MHK7HWWIBHTA1HVKSCY45WZ_?usp=drive_link)

LINK GITHUB: [HTTPS://GITHUB.COM/WALKERVZ/FINAL-PROJECT-ID-X-PARTNERS/](https://github.com/walkervz/final-project-id-x-partners/)