



# FINAL TASK HOME CREDIT

Prediksi Risiko Gagal Bayar  
Nasabah Home Credit  
Menggunakan Machine Learning

2025

AURA AULIA ANASTASYA HADI PUTRI



LINKEDIN





# LATAR BELAKANG

Home Credit menghadapi tantangan dalam memastikan bahwa pelanggan yang layak tidak ditolak saat mengajukan pinjaman. Model prediksi risiko yang akurat diperlukan agar keputusan kredit dapat lebih tepat sasaran.

Rumusan Masalah : Bagaimana mengidentifikasi pelanggan berisiko tinggi gagal bayar?

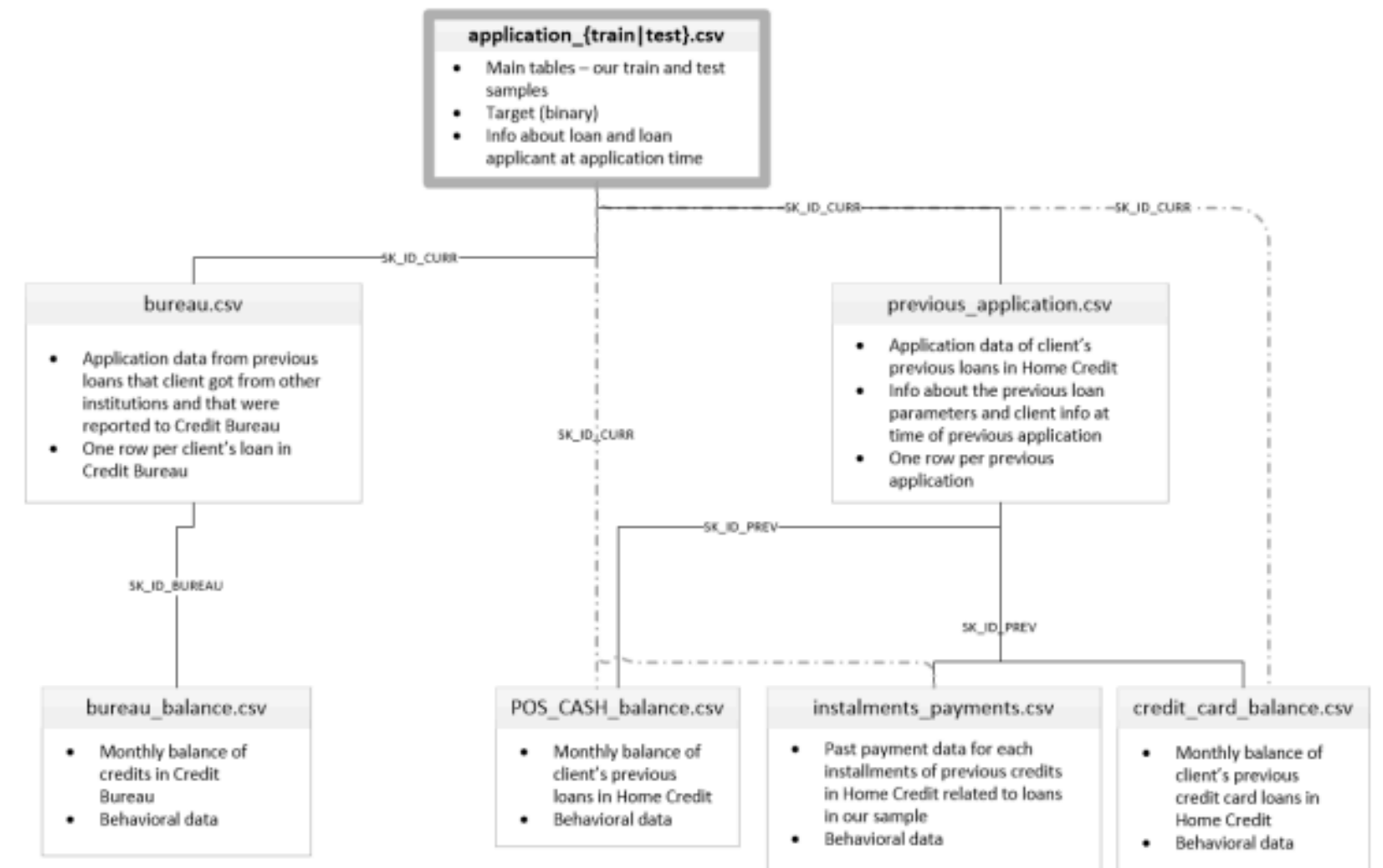
Tujuan : Memprediksi risiko gagal bayar nasabah baru berdasarkan data historis.

Tantangan : Data yang kompleks dan imbalanced, namun dapat dilakukan eksperimen hingga menemukan model yang sesuai.



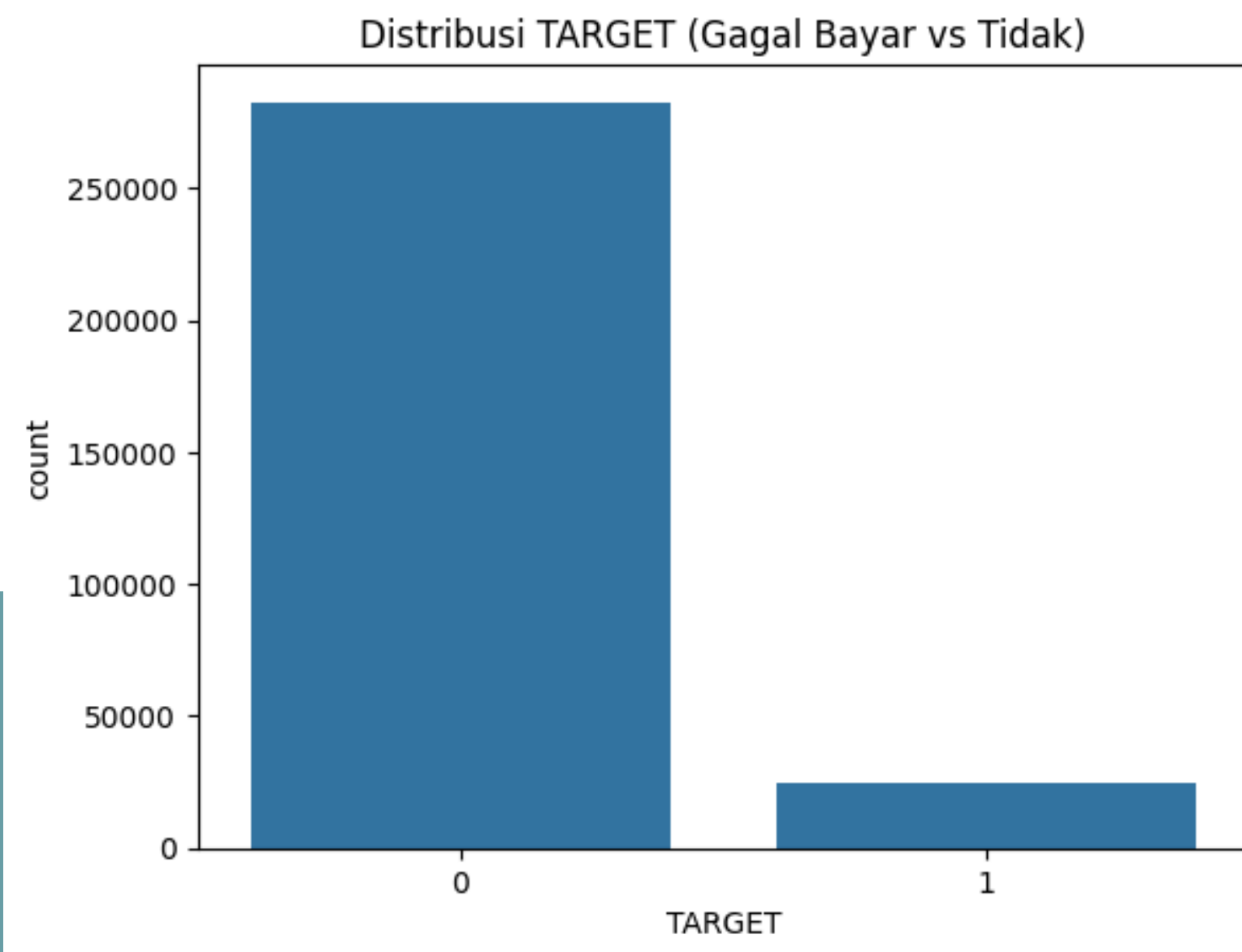
# SUMBER DATASET

Proyek ini menggunakan lebih dari 7 dataset berbeda, termasuk application\_train.csv, bureau.csv, previous\_application.csv, credit\_card\_balance.csv, POS\_CASH\_balance.csv, dan lainnya. Semua tabel tersebut digabungkan berdasarkan SK\_ID\_CURR untuk membentuk satu dataset utama. Kemudian dilakukan proses feature engineering untuk menghasilkan fitur-fitur seperti rasio pendapatan terhadap kredit (CREDIT\_INCOME\_RATIO), usia nasabah dari DAYS\_BIRTH (AGE), jumlah pengajuan sebelumnya yang ditolak (NUM\_PREV\_REFUSED), dan lainnya. Dataset akhir memiliki 139 kolom dan lebih dari 300.000 baris data yang kemudian digunakan sebagai input dalam pemodelan.

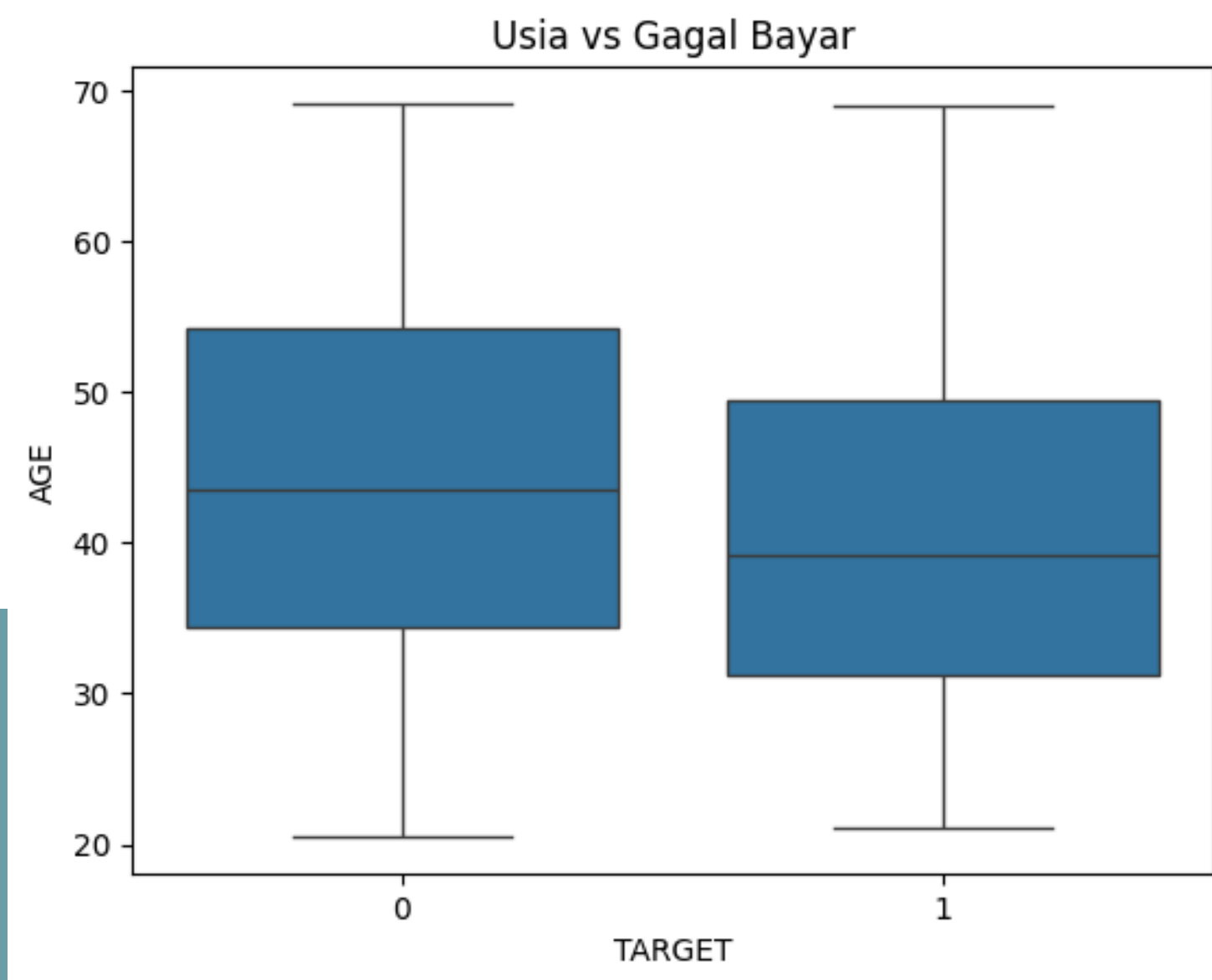




# DATA EXPLORATION



- Mayoritas data menunjukkan tidak gagal bayar (TARGET = 0), jauh lebih banyak dibandingkan yang gagal bayar (TARGET = 1)
- Dataset imbalance:



- Rata-rata usia nasabah yang gagal bayar (TARGET = 1) lebih rendah dibandingkan yang tidak gagal bayar (TARGET = 0.)
- Penyebaran usia nasabah yang gagal bayar lebih sempit. Hal ini mengindikasikan bahwa nasabah yang lebih muda cenderung lebih berisiko untuk gagal bayar dibandingkan yang lebih tua.



# LOGISTIC REGRESSION

- Akurasi model sebesar 62%, yang berarti 62% prediksi model sesuai dengan label sebenarnya. Namun, karena data tidak seimbang (jumlah tidak gagal bayar jauh lebih banyak), akurasi bukanlah metrik yang sepenuhnya merepresentasikan performa model.
- Kinerja pada kelas mayoritas (Tidak Gagal Bayar - label 0):
- Precision tinggi (0.95), menunjukkan bahwa sebagian besar prediksi "tidak gagal bayar" memang benar.
- Namun, recall rendah (0.62), artinya model sering gagal mengenali semua kasus "tidak gagal bayar".
- Kinerja pada kelas minoritas (Gagal Bayar - label 1):
- Precision sangat rendah (0.12), artinya banyak prediksi "gagal bayar" yang salah (false positive).
- Recall cukup tinggi (0.60), artinya model masih cukup baik dalam mengenali sebagian besar kasus gagal bayar.
- F1-score (0.20) rendah, menunjukkan ketidakseimbangan antara precision dan recall.
- Nilai ROC AUC Score sebesar 0.645, yang mengindikasikan bahwa kemampuan model dalam membedakan antara kelas gagal bayar dan tidak gagal bayar masih terbatas, namun lebih baik dibanding tebakan acak (0.5).

## Logistic Regression

```
[[34860 21678]
```

```
[ 1983  2982]]
```

	precision	recall	f1-score	support
0	0.95	0.62	0.75	56538
1	0.12	0.60	0.20	4965
accuracy			0.62	61503
macro avg	0.53	0.61	0.47	61503
weighted avg	0.88	0.62	0.70	61503

ROC AUC Score: 0.6452883973231276

Kesimpulan : Model mampu mengenali sebagian besar kasus gagal bayar (recall cukup baik), namun sering salah dalam mengklasifikasikan data sebagai gagal bayar (precision rendah). Hal ini menunjukkan perlunya perbaikan model terutama dalam menangani ketidakseimbangan kelas, misalnya dengan resampling, penyesuaian threshold, atau menggunakan model lain seperti Random Forest atau XGBoost.



# XGBOOST

	precision	recall	f1-score	support
0	0.94	0.77	0.85	56538
1	0.15	0.48	0.23	4965
accuracy			0.74	61503
macro avg	0.55	0.62	0.54	61503
weighted avg	0.88	0.74	0.80	61503

Kesimpulan : Meskipun menghasilkan akurasi yang bagus, perlu dilakukan peningkatan kembali.

Model XGBoost menunjukkan hasil yang lebih baik dibandingkan model sebelumnya. Akurasi model mencapai 74%, yang berarti sebagian besar prediksi sudah sesuai dengan label sebenarnya.

Pada kelas tidak gagal bayar (label 0), model memiliki performa yang cukup kuat dengan precision 94% dan recall 77%, mencerminkan bahwa model sangat baik dalam mengenali dan mengklasifikasikan nasabah yang membayar dengan lancar.

Namun, pada kelas gagal bayar (label 1), model masih menghadapi tantangan. Precision-nya sebesar 15%, artinya dari seluruh prediksi gagal bayar, hanya sebagian kecil yang benar-benar sesuai. Meskipun begitu, recall-nya mencapai 48%, yang berarti hampir setengah dari seluruh nasabah yang benar-benar gagal bayar berhasil terdeteksi oleh model.

Nilai F1-score untuk kelas gagal bayar sebesar 0.23, menunjukkan adanya keseimbangan awal antara kemampuan model dalam mendeteksi dan menghindari prediksi yang salah, meski masih perlu ditingkatkan.





# XGBOOST TUNED

```
[[38142 18396]
 [ 1964  3001]]
      precision    recall  f1-score   support

     0       0.95      0.67      0.79     56538
     1       0.14      0.60      0.23      4965

 accuracy          0.67     61503
 macro avg       0.55      0.64      0.51     61503
 weighted avg    0.89      0.67      0.74     61503

ROC AUC Score: 0.6895654953808927
```

## Kesimpulan :

Model ini dipilih karena mampu menjaga keseimbangan antara akurasi dan deteksi risiko. Meskipun akurasinya tidak tertinggi, recall untuk kasus gagal bayar cukup tinggi (60%), yang penting dalam konteks mitigasi risiko. Threshold disesuaikan untuk hasil yang lebih stabil dan sesuai kebutuhan analisis.

Model ini merupakan hasil dari optimasi hyperparameter menggunakan RandomizedSearchCV dengan 20 kombinasi parameter, serta penyesuaian threshold menggunakan ROC Curve, di mana nilai optimal ditemukan pada threshold 0.47.

Berdasarkan hasil evaluasi, model mencapai akurasi sebesar 67%, dengan recall untuk kelas gagal bayar (TARGET = 1) sebesar 60%, dan nilai ROC AUC sebesar 0.689. Artinya, model memiliki kemampuan yang cukup baik dalam membedakan antara nasabah yang berisiko gagal bayar dan yang tidak.

Confusion matrix menunjukkan bahwa dari total 4.965 nasabah yang benar-benar gagal bayar, model berhasil mengidentifikasi 3.001 di antaranya. Meskipun precision untuk kelas ini masih rendah (14%), recall yang tinggi menjadi nilai tambah, terutama karena dalam konteks risiko, lebih baik menangkap sebanyak mungkin potensi gagal bayar meskipun terdapat false positive.





# REKOMENDASI BISNIS

2025

Berdasarkan hasil model yang telah dikembangkan, disarankan agar skor prediksi risiko yang dihasilkan oleh model digunakan secara aktif dalam proses evaluasi pengajuan kredit. Skor ini dapat dimanfaatkan untuk memberikan peringatan otomatis saat sistem mendeteksi nasabah dengan probabilitas gagal bayar yang tinggi. Selain itu, skor ini dapat dijadikan dasar dalam menyesuaikan struktur produk kredit, seperti menentukan limit pinjaman dan tenor pembayaran yang lebih sesuai dengan profil risiko nasabah.

Nasabah dengan karakteristik tertentu, seperti usia muda dan riwayat pengajuan pinjaman yang sering ditolak, sebaiknya ditempatkan dalam jalur verifikasi tambahan untuk memastikan validitas data dan kemampuan bayar mereka. Dengan integrasi skor prediksi ini ke dalam workflow approval kredit, diharapkan proses persetujuan menjadi lebih akurat dan efisien, sekaligus mengurangi potensi kerugian akibat gagal bayar.





# PENUTUP

Proyek ini merupakan implementasi end-to-end dalam memprediksi risiko gagal bayar nasabah menggunakan pendekatan Machine Learning. Dimulai dari pemahaman masalah dan eksplorasi data, dilanjutkan dengan penggabungan berbagai sumber dataset, pemrosesan fitur, hingga pemodelan menggunakan Logistic Regression dan XGBoost. Model terbaik yang dipilih adalah XGBoost hasil tuning, dengan threshold optimal 0.47 yang menghasilkan recall TARGET=1 sebesar 60% dan ROC AUC sebesar 0.689. Model ini mampu memberikan prediksi yang seimbang dan dapat diandalkan dalam mendeteksi calon nasabah berisiko tinggi. Seluruh proses dan kode proyek ini dapat diakses melalui link berikut:

Terima kasih atas perhatian dan kesempatan yang diberikan.

[GitHub](#)

[Drive](#)

[LinkedIn](#)

[Portfolio](#)