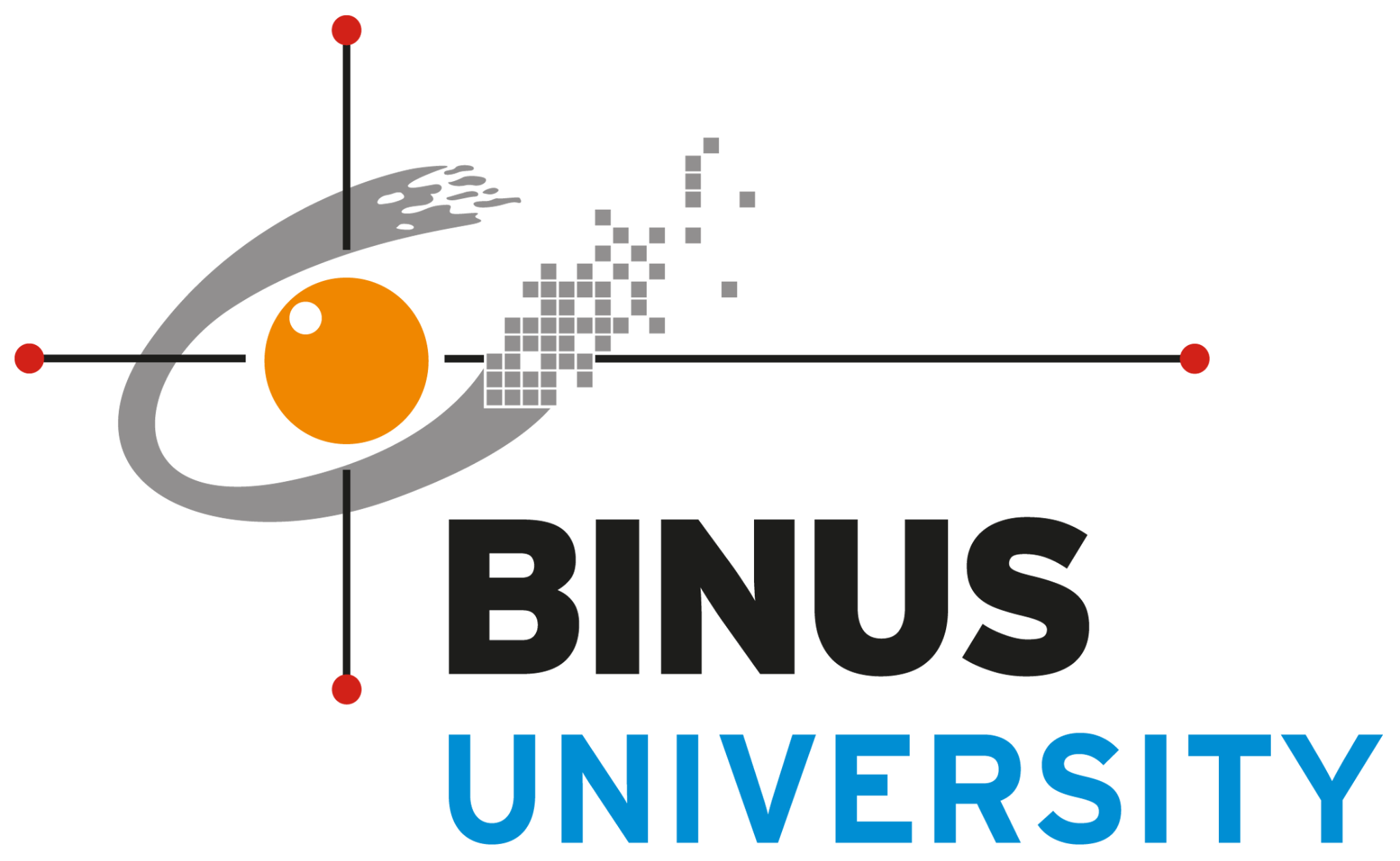
**Credit Rate Analysis**



**Disusun oleh:**

Frenky / 2602102302

Haikal Syafi Muflih / 2602106446

Tatiya Sucitta / 2602110323

**DAFTAR ISI**

[I. PENDAHULUAN 2](#_r7zpx3yfrkao)

[II. STUDI LITERATUR 3](#_oitg5pouguwp)

[III. METODE 4](#_h7bpv3qhi96q)

[IV. HASIL IMPLEMENTASI 5](#_26i31726g5l2)

[V. KESIMPULAN 10](#_5vblsllsq763)

[Referensi 11](#_9o1y5on08n3)

[Lampiran 11](#_wwskpgw4g5uy)

# **PENDAHULUAN**

Dalam dunia pasar keuangan yang terus berubah, penilaian kredit menjadi sangat penting bagi investor dan peminjam. Dengan kemajuan instrumen keuangan dan platform peminjaman, pemahaman tentang risiko kredit semakin kompleks. Evaluasi yang akurat terhadap kelayakan kredit tidak hanya melindungi kepentingan investor, tetapi juga memastikan penggunaan modal yang efisien dalam perekonomian. Oleh karena itu, tulisan ini bertujuan untuk menjelajahi lebih dalam tentang penilaian kredit, dengan maksud memberikan pemahaman tentang metode dan kerangka kerja yang digunakan untuk mengevaluasi kemungkinan peminjam memenuhi kewajiban keuangan mereka, terutama dalam konteks pinjaman yang melibatkan bunga.

Penilaian kredit, sebagai elemen kunci dalam dunia keuangan modern, menjadi instrumen penting bagi investor dan lembaga keuangan untuk mengevaluasi risiko yang terkait dengan kegiatan peminjaman. Metrik konvensional seperti rasio keuangan, perbandingan industri, dan faktor kualitatif telah diolah dengan pengembangan *Machine Learning* dan analisis *Big Data* untuk meningkatkan ketepatan prediksi.

Tulisan ini bertujuan untuk menggali lebih dalam tentang kompleksitas penilaian kredit, memungkinkan penggunaan modal yang efisien dan memberikan pengetahuan yang bermanfaat bagi investor dan pemberi pinjaman agar dapat membuat keputusan yang tepat di tengah dinamika lanskap keuangan.

# **STUDI LITERATUR**

**Prj3\_loan credit prediction, loan-project**

* *Logistic regression* teknik analisis data yang menggunakan matematika untuk menemukan hubungan antara dua faktor data. Hubungan ini kemudian digunakan untuk memprediksi nilai salah satu faktor berdasarkan faktor lainnya. Prediksi biasanya mempunyai jumlah hasil yang terbatas, seperti ya atau tidak.
* *Random Forest* adalah *machine learning* yang digunakan untuk tugas klasifikasi dan regresi. Ini termasuk dalam kelompok pembelajaran ansambel, yang menggabungkan beberapa model individual untuk menghasilkan model prediksi yang lebih akurat dan kuat.
* XGBoost (eXtreme Gradient Boosting) adalah algoritma pembelajaran mesin yang membuat model prediksi dengan menggabungkan prediksi dari beberapa model, dan menggunakan *gradient boosted decision tree* (GDBT)

**Loan Credit Prediction**

* *Random forest* dengan *Tuned Random Forest*, dimana dilakukan pemilihan parameter yang terbaik dari grid yang sudah ditentukan sebelumnya dengan menggunakan RandomizedSearchCV(). Syntax ini menggunakan kombinasi dari parameter secara acak hingga dapat menemukan pasangan parameter yang menghasilkan score terbaik.

# **METODE**

Dalam prediksi *credit rate*, beberapa metode populer yang sering digunakan adalah Logistic regression, random forest, dan XGBoost. Logistic regression adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel independen dan variabel dependen biner, yang cocok untuk memprediksi apakah seorang peminjam akan gagal atau berhasil dalam membayar pinjamannya. Random forest adalah metode ensemble yang terdiri dari banyak decision tree, di mana setiap decision tree dihasilkan secara acak dan hasilnya diambil secara agregat untuk memberikan prediksi yang lebih stabil dan akurat.

Selain itu, Logistic Regression juga dapat menjadi salah satu pertimbangan model yang bisa diterapkan. Pada Logistic Regression, data pinjaman dikumpulkan dan dipersiapkan untuk analisis. Ini mencakup pembersihan data, penghapusan nilai yang hilang, dan pengkodean variabel kategori menjadi bentuk yang dapat dimengerti oleh model. Selanjutnya, variabel yang berpotensi mempengaruhi prediksi (seperti pendapatan, skor kredit, jumlah pinjaman, dll.) dipilih dan dimasukkan ke dalam model Logistic Regression. Model kemudian dilatih menggunakan data pelatihan, di mana parameter model disesuaikan atau *fitted* untuk mencocokkan data yang nantinya akan diamati.

Penggunaan model Logistic Regressionyang optimal pada kasus ini tergantung pada berbagai faktor seperti kompleksitas data, ukuran sampel, dan tujuan prediksi. Namun, dalam banyak kasus, model ini sering dianggap terlalu sederhana. Oleh karena itu, metode yang lebih canggih seperti *ensemble learning* (*Random Forest* atau *XGBoost*) atau *deep learning* mungkin memberikan hasil yang lebih baik karena kemampuannya untuk menangani pola dan interaksi yang kompleks antara atribut di dataset.

Namun pada kasus dataset *Loan Data*, atribut yang dimiliki bisa dibilang cenderung sederhana dan singkat. Meskipun begitu, kami telah membuat perbandingan dari tiga buah model yakni Logistic Regression, Random Forest, dan XGBoost. Menurut pembacaan referensi kami, metode yang lebih kami rekomendasikan adalah Logistic Regression.

# **HASIL IMPLEMENTASI**

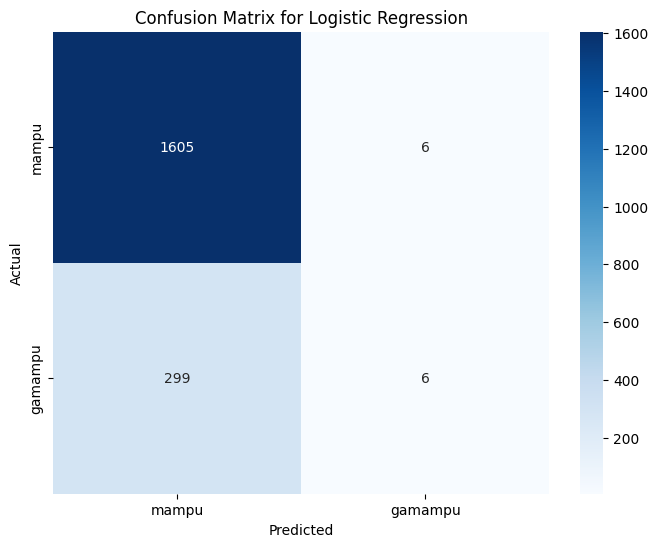
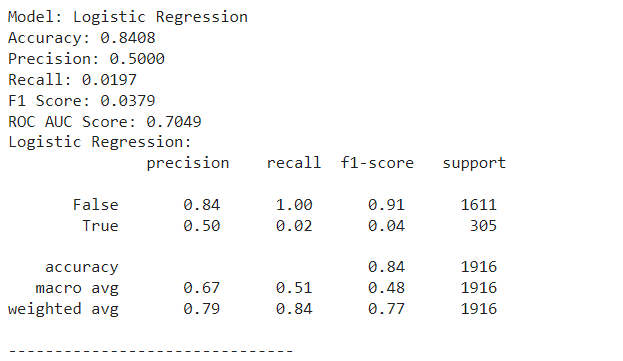
Pada kasus kali ini, tingkat akurasi dan *false negative* adalah kedua parameter yang paling diperhatikan dalam pemilihan model terbaik.Output yang dihasilkan oleh model sendiri akan berupa *boolean* dimana 1 berarti tidak mampu membayar pinjaman dan 0 berarti mampu membayar pinjaman.

Akurasi sendiri adalah kemampuan model dalam mengklasifikasikan *input* dari pengguna dengan tepat, baik itu positif maupun negatif. Ini adalah pengukuran palling umum yang digunakan untuk mengevaluasi hasil kerja model.

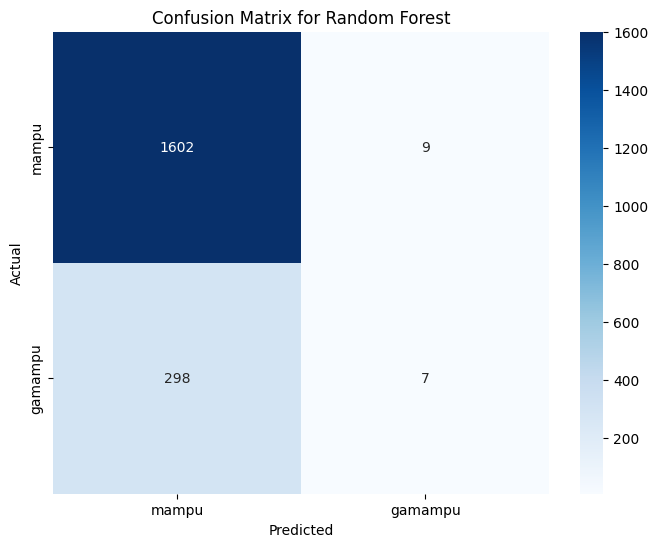
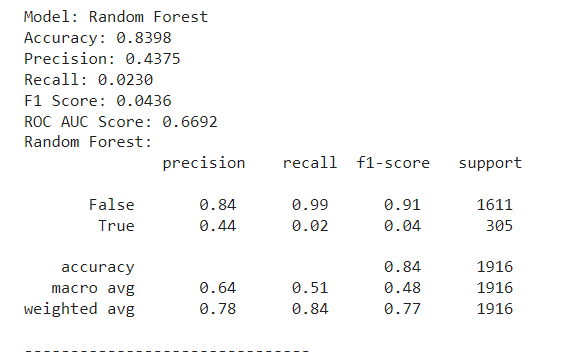
Melainkan *false negative* merupakan salah satu parameter yang berada di *confusion matrix* yang berarti jumlah sampel positif yang salah diprediksi sebagai negatif oleh model (label aktual “gamampu”, label diprediksi “mampu”). Singkatnya ini berarti model salah menganggap bahwa sebuah instansi mampu membayar pinjaman walaupun sebenarnya tidak mampu.

Pengimplementasian pada code yang telah kami susun menunjukkan bahwa pada tahapan prediksi yang pertama, hasil prediksi ketiganya menunjukkan data sebagai berikut:

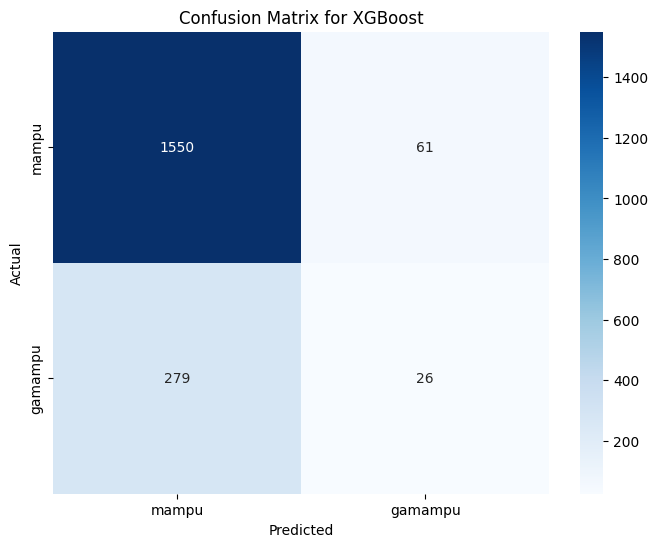
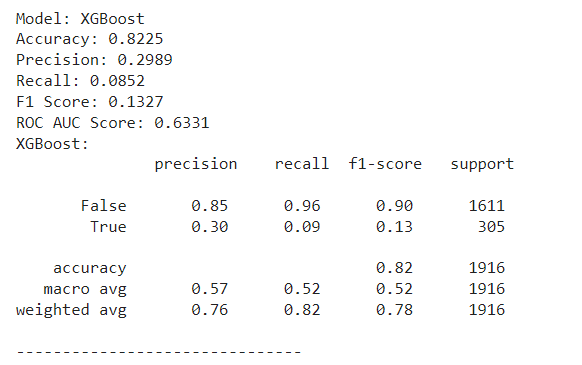
* **Logistic Regression**



* **Random Forest**



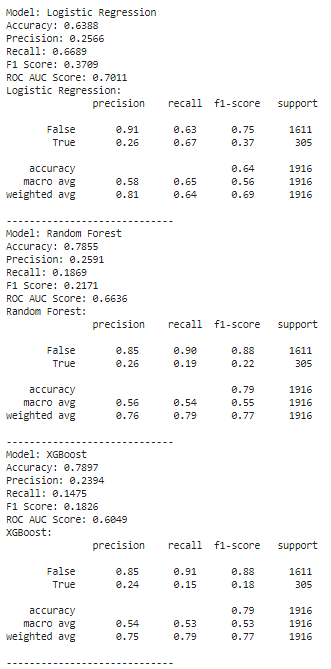
* **XGBoost**

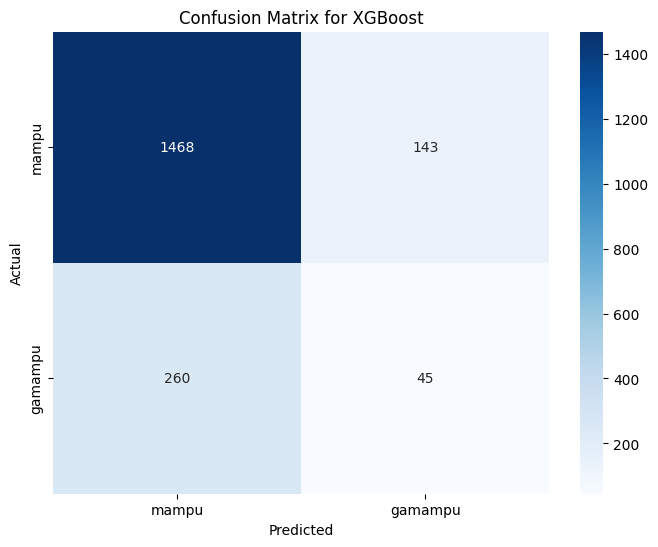
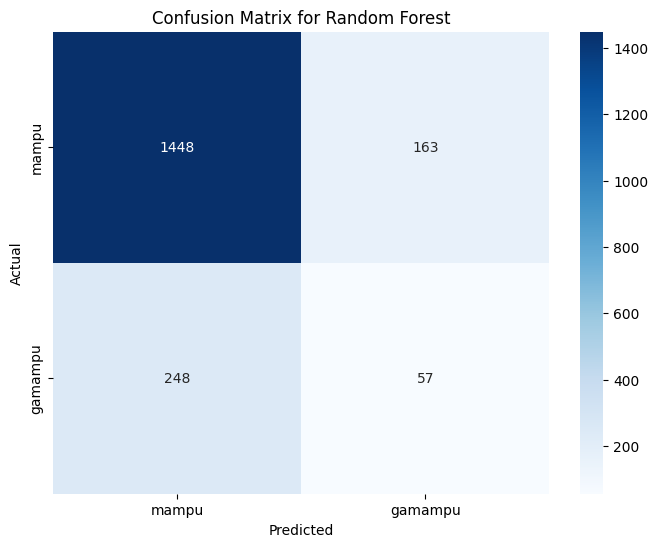
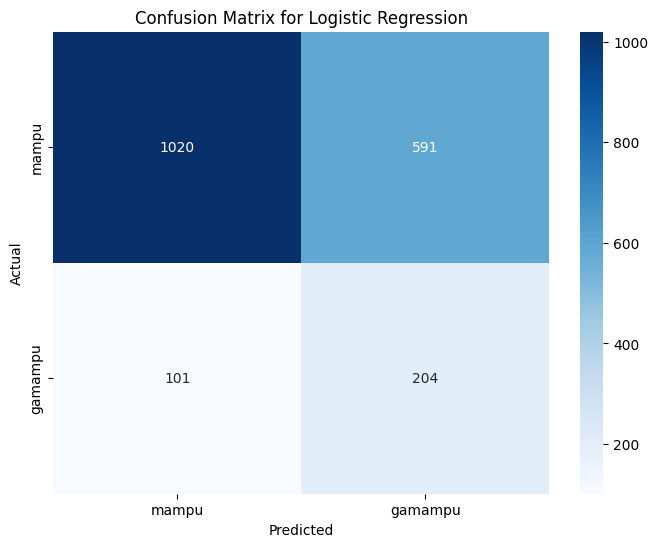
**

Pada tahap pertama, Logistic Regression danRandom Forest memiliki akurasi yang lebih tinggi 2% dari model XGBoost. Disamping itu, XGBoost ternyata memiliki tingkat resistansi atas *false negative* yang paling rendah.

Selanjutnya pada tahap kedua yaitu setelah dilakukannya SMOTE. SMOTE merupakan library bawaan dari python yang memungkinkan untuk mengatasi masalah *imbalanced data* dengan melakukan sintesis data dari kelas minoritas (pada kasus ini instansi yang tidak mampu membayar pinjaman) dengan pembentukan *convex* kombinasi dari instances yang saling berdekatan di dataset tanpa mengakibatkan *overfitting*.

Prediksi yang dihasilkan sebagai berikut:





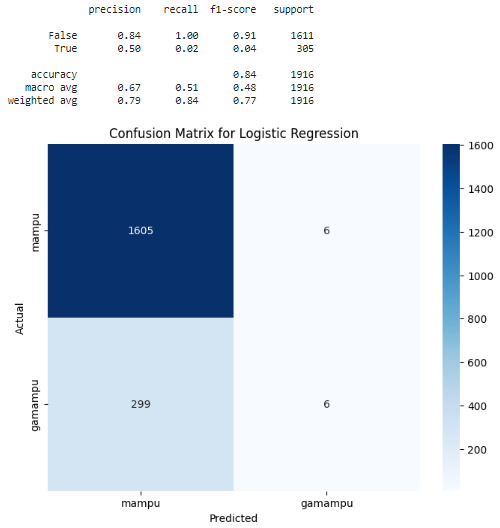
Pada tahap kedua, terlihat penurunan nilai akurasi Logistic Regressionsebanyak 20% yang mengakibatkan model Random Forest dan XGBoost unggul dengan nilai keduanya 79%. Meskipun kedua model lainnya terlihat seimbang, tetapi sebenarnya model yang mengalami penurunan akurasi paling minimum ialah XGBoost dengan hanya 3% sedangkan Random Forest 5%.

Meskipun begitu, nilai *false negative* terendah lagi-lagi dimiliki oleh model dengan akurasi terendah, yaitu *Logistic Regression.*

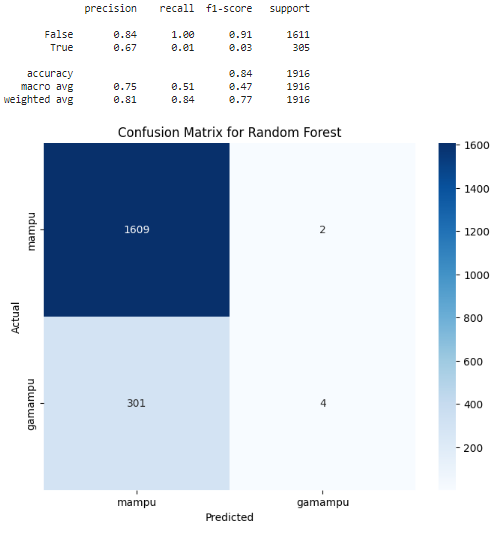
tahapan terakhir atau ketiga adalah dengan menggunakan library RandomizedSearchCV dimana dapat memberikan set *hyperparameter* yang nantinya dapat kita gunakan untuk menghasilkan output terbaik. Parameter-parameter akan diberikan dan RandomizedSearchCV akan melakukan pengacakan untuk cara mencari kombinasi *hyperparameter* yang terbaik.

Berikut adalah hasil prediksinya:

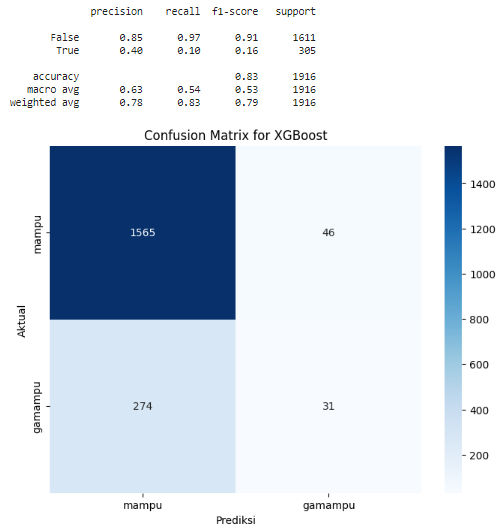
* **Logistic Regression**



* **Random Forest**



* **XGBoost**



Setelah menggunakan parameter yang dianggap terbaik dari RandomizedSearchCV, dihasilkan *Logistic Regression* yang memiliki skor dan matriks yang sama seperti pada tahap pertama. Random Forest juga mendapatkan skor akurasi yang sama seperti pada tahapan pertama namun dengan akumulasi *confusion matrix* yang berbeda seperti peningkatan jumlah *false negative*.

Terakhir, XGBoost menempati skor akurasi tertinggi dari hasil XGBoost di kedua tahap sebelumnya dengan skor 83%. Menariknya, selain akurasi yang paling tinggi, model ini juga mendapatkan *false negative* yang paling rendah diantara 2 model lainnya di tahap terakhir dengan nilai 274.

# **KESIMPULAN**

Pada tahap prediksi pertama, model Logistic Regression dan Random Forest telah lebih mengungguli XGBoost dengan score keduanya 84% sama. Namun, XGBoost memiliki nilai false negative terkecil diantara kedua model lainnya.

Setelah itu, pada tahapan setelah SMOTE atau setelah pensintesisan data minoritas, tingkat keakuratan yang dihasilkan oleh Logistic Regression memang belum bisa mengungguli model dari Random Forest dan XGBoost.

Diantara seluruh proses dan analisis, Logistic Regression pada tahap setelah SMOTE merupakan model yang memiliki score *false negative* paling rendah. *False negative* pada kasus kami ini berarti **instansi yang kemungkinan besar tidak mampu membayar pinjaman**, salah diprediksi oleh mesin sebagai **instansi yang mampu mengembalikan pinjamannya**. Disamping itu, dapat dilihat pada setiap hasil confusion matrix dengan nilai false negative rendah selalu memiliki tingkat akurasi paling rendah juga. Ini membuktikan bahwa akurasi sendiri tidak selalu mencerminkan kualitas model yang lebih baik, melainkan faktor lain juga perlu dipertimbangkan seperti confusion matrix yang ternyata dapat menyajikan data yang lebih rinci tentang kualitas model baik dalam menangani jumlah data minoritas maupun jumlah kesalahan prediksi yang dilakukan oleh model.

Maka dari itu, kami lebih mengutamakan peminimalisiran kerugian pengguna (para investor dan bank) dengan mencegah pengguna salah memberikan modal kepada pihak yang kurang menjanjikan seperti pencegahan score false negative yang tinggi.

# **Referensi**

1. Girsang, A. S. (2018, June 8). *Synthetic Minority Over-sampling Technique (SMOTE) algorithm for handling imbalanced data*. Master of Computer Science. <https://mti.binus.ac.id/2018/06/08/synthetic-minority-over-sampling-technique-smote-algorithm-for-handling-imbalanced-data/>
2. *RandomizedSearchCV to find Optimal Parameters in Python -*. (2023, January 19). ProjectPro. <https://www.projectpro.io/recipes/find-optimal-parameters-using-randomizedsearchcv-for-regression>
3. *What is XGBoost?* (n.d.). NVIDIA Data Science Glossary. <https://www.nvidia.com/en-us/glossary/xgboost/>
4. Simplilearn. (2023, November 7). *What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning*. Simplilearn.com. <https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article#:~:text=cost%20to%20accuracy!-,What%20is%20XGBoost%20Algorithm%3F,optimize%20their%20machine%2Dlearning%20models>.
5. Guest\_Blog. (2024, January 16). *XGBOOST: Introduction to XGBOOST algorithm in Machine Learning*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/>
6. S, P. (2024, February 7). *Logistic Regression Model: A guide to machine learning techniques and applications*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regression-model/>
7. Ridhijhamb. (2021, September 29). *Loan prediction| Random Forest*. Kaggle. <https://www.kaggle.com/code/ridhijhamb/loan-prediction-random-forest>
8. Aidaghanbari. (2024, February 3). *Loan-Project*. Kaggle. <https://www.kaggle.com/code/aidaghanbari/loan-project>
9. Maseratiurm. (2023, December 21). *Prj3\_loan credit prediction*. Kaggle. <https://www.kaggle.com/code/maseratiurm/prj3-loan-credit-prediction>
10. *Loan data*. (2021, May 14). Kaggle. <https://www.kaggle.com/datasets/itssuru/loan-data?resource=download>

# **Lampiran**