**Étapes pour le Traitement des Données**

1. Chargement des Données: Charger les fichiers CSV (lovoo\_v3\_users\_api-results.csv et lovoo\_v3\_users\_instances.csv). La fonction pd.read\_csv charge les fichiers CSV dans des objets DataFrame.
2. Suppression des Colonnes Inutiles: Identifier et supprimer les colonnes non nécessaires à l'analyse. Sélectionne uniquement les colonnes définies dans columns\_to\_keep.
3. Filtrage des Comptes Bannis: Exclure toutes les lignes où le compte est marqué comme true dans la colonne locked
4. Gestion des Données Duplicates: Déterminer les critères pour choisir quelle entrée conserver en cas de duplication. Par exemple, vous pourriez préférer les valeurs provenant de api\_data ou fusionner les informations des deux datasets.
5. Fusionner les Données : Fusionner les datasets sur userId.

L'objectif de cette analyse est de déterminer la popularité des profils sur un site de rencontre en utilisant des critères tels que le nombre de visites et le taux de "kisses". Les colonnes sélectionnées sont choisies pour leur pertinence directe à cette analyse. Voici une explication détaillée de chaque colonne conservée et la raison pour laquelle d'autres colonnes ne sont pas incluses.

**Colonnes Sélectionnées et Leur Importance**

userId : Identifiant unique de l'utilisateur, essentiel pour identifier chaque profil de manière unique et pour gérer les doublons lors de la fusion des datasets.

name : Nom d'utilisateur. Bien que non directement utilisé dans l'analyse de popularité, il peut être utile pour des références visuelles et pour valider les données.

age : Âge de l'utilisateur. C'est une caractéristique démographique importante qui pourrait influencer la popularité d'un profil. Les profils plus jeunes peuvent avoir des dynamiques différentes par rapport aux profils plus âgés.

city : Ville de l'utilisateur. La localisation géographique peut jouer un rôle significatif dans la popularité des profils, par exemple, les profils dans les grandes villes pourraient avoir des dynamiques différentes par rapport aux profils dans les petites villes.

counts\_profileVisits : Nombre de visites de profil. C'est une métrique directe de la popularité d'un profil. Plus le nombre de visites est élevé, plus le profil est potentiellement populaire.

counts\_kisses : Nombre de "kisses" reçus. C'est une autre métrique directe de l'engagement et de la popularité d'un profil. Un taux élevé de "kisses" peut indiquer un profil attractif.

counts\_fans : Nombre de fans. Cela montre combien de personnes suivent ou sont intéressées par le profil, ce qui est un autre indicateur de popularité.

counts\_pictures : Nombre de photos. Les profils avec plus de photos peuvent attirer plus de visites et d'interactions, ce qui est pertinent pour mesurer la popularité.

flirtInterests\_chat : Intérêt pour le chat. Cela indique si l'utilisateur est intéressé par le chat, ce qui peut influencer le nombre de visites et de "kisses" qu'un profil reçoit. Les utilisateurs intéressés par le chat peuvent recevoir plus d'interactions.

flirtInterests\_friends : Intérêt pour faire des amis. Cela pourrait également influencer la popularité, car les utilisateurs cherchant à élargir leur cercle social peuvent recevoir plus de visites.

flirtInterests\_date : Intérêt pour les rendez-vous. De même, cela peut être un facteur clé de popularité pour ceux cherchant activement des rencontres.

Colonnes Non Sélectionnées et Raisons

connectedToFacebook, isVIP, isVerified, isFlirtstar, isHighlighted, isInfluencer : Bien que ces colonnes puissent indiquer un certain niveau de statut ou de vérification, elles ne sont pas directement liées aux critères de popularité définis (visites et "kisses"). Ces informations peuvent être pertinentes pour d'autres types d'analyses, mais elles ne sont pas essentielles ici.

locationCity, locationCitySub : Détailler davantage la localisation n'apporte pas de valeur supplémentaire significative pour notre analyse principale qui se concentre sur la popularité.

counts\_g : Bien qu'ils puissent indiquer une certaine forme d'engagement, ils ne sont pas aussi directement liés aux critères définis pour la popularité (visites et "kisses").

birthd, crypt, freetext, whazzup : Ces informations personnelles et textuelles ne sont pas nécessaires pour l'analyse quantitative de la popularité.

lastOnlineDate, lastOnlineTime, distance : Bien que ces colonnes puissent fournir des informations supplémentaires sur l'activité de l'utilisateur et la proximité, elles ne sont pas directement pertinentes pour l'analyse de popularité basée sur les visites et les "kisses".

**Étapes pour appliquer Random Forest :**

1. Préparation des Données :
   * Transformer les colonnes catégorielles en variables numériques.
   * Gérer les valeurs manquantes.
   * Diviser les données en ensembles d'entraînement et de test.
2. Entraînement du Modèle :
   * Entraîner un modèle Random Forest sur les données d'entraînement.
3. Évaluation du Modèle :
   * Évaluer les performances du modèle sur l'ensemble de test

**Explication des Étapes pour l'Optimisation du Modèle**

1. **Définir les Hyperparamètres** :
   * n\_estimators : Nombre d'arbres dans la forêt.
   * max\_features : Nombre de caractéristiques à considérer pour chaque division.
   * max\_depth : Profondeur maximale de chaque arbre.
   * criterion : Fonction pour mesurer la qualité d'une division (Gini ou Entropie).
2. **Initialiser GridSearchCV** :
   * param\_grid : Grille de paramètres à tester.
   * cv : Nombre de plis pour la validation croisée.
   * scoring : Métrique à optimiser (ici, accuracy).
3. **Exécuter GridSearchCV** :
   * fit : Entraîner GridSearchCV sur les données d'entraînement.
4. **Récupérer les Meilleurs Hyperparamètres** :
   * best\_params\_ : Les meilleurs hyperparamètres trouvés.
   * best\_estimator\_ : Le meilleur modèle entraîné avec ces hyperparamètres.
5. **Évaluation du Modèle Optimisé** :
   * accuracy\_score et classification\_report pour évaluer les performances du modèle optimisé sur l'ensemble de test.

Meilleurs Hyperparamètres

Meilleurs hyperparamètres : {'criterion': 'gini', 'max\_depth': 20, 'max\_features': 'sqrt', 'n\_estimators': 200}

Accuracy après optimisation: 0.9429097605893186

Classification Report après optimisation:

precision recall f1-score support

0 0.95 0.97 0.96 397

1 0.93 0.86 0.89 146

accuracy 0.94 543

macro avg 0.94 0.92 0.93 543

weighted avg 0.94 0.94 0.94 543

**Interprétation des Résultats Après Optimisation**

Précision Globale (Accuracy) :

La précision globale du modèle est passée de 93.74% à 94.29%, ce qui indique une légère amélioration.

Classe 0 (Non populaires) :

Précision : 95% (contre 94% avant optimisation).

Rappel : 97% (contre 98% avant optimisation).

F1-Score : 96% (inchangé).

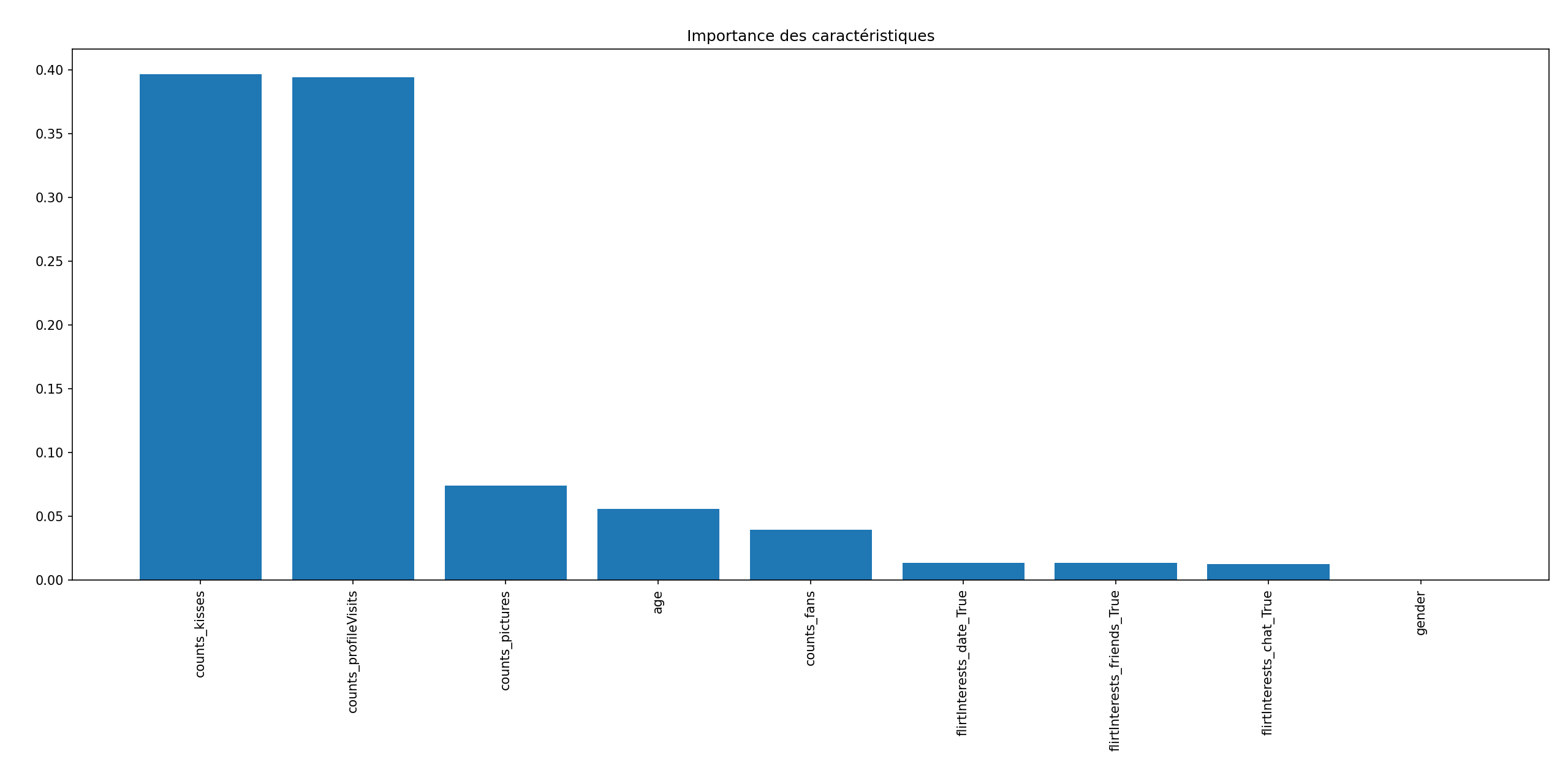
Classe 1 (Populaires) :

Précision : 93% (contre 94% avant optimisation).

Rappel : 86% (contre 82% avant optimisation).

F1-Score : 89% (contre 88% avant optimisation).

**Construction d’importance des carateristiques**

 Le graphique montre l'importance relative des différentes caractéristiques utilisées dans le modèle Random Forest optimisé. Voici une interprétation des résultats :

counts\_kisses :

C'est la caractéristique la plus importante, indiquant que le nombre de "kisses" reçus joue un rôle crucial dans la prédiction de la popularité d'un profil.

counts\_profileVisits :

Cette caractéristique est presque aussi importante que counts\_kisses, suggérant que le nombre de visites de profil est également un indicateur clé de la popularité.

counts\_pictures :

Le nombre de photos contribue significativement à la prédiction, ce qui peut indiquer que les profils avec plus de photos sont plus susceptibles d'être populaires.

age :

L'âge a une importance modérée, ce qui peut refléter des préférences d'âge spécifiques parmi les utilisateurs du site.

counts\_fans :

Le nombre de fans joue également un rôle, bien que moins significatif que les visites de profil et les "kisses".

flirtInterests\_date\_True :

L'intérêt pour les rendez-vous a une certaine importance, ce qui montre que les utilisateurs cherchant activement des rencontres peuvent influencer la popularité.

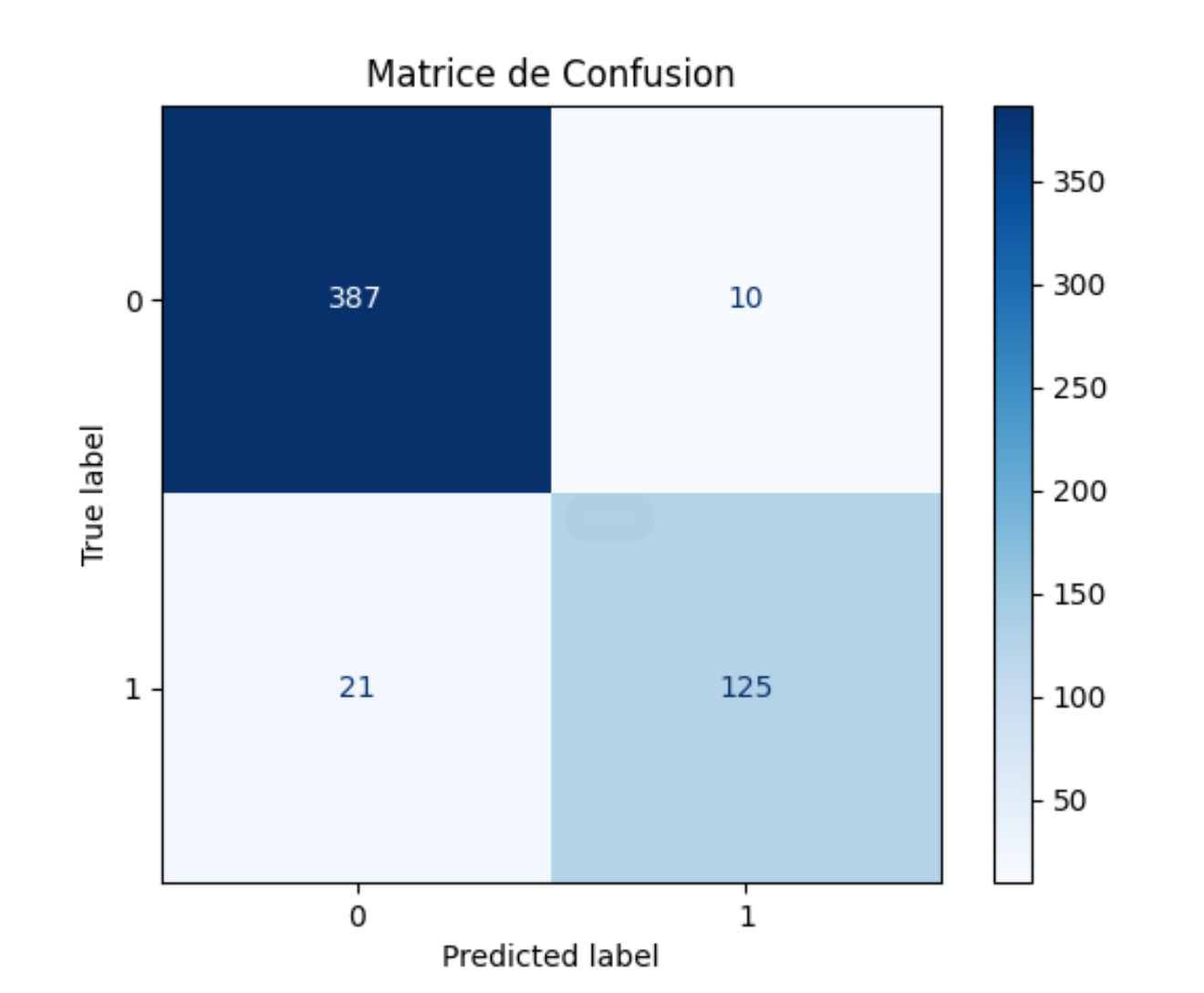
flirtInterests\_friends\_True et flirtInterests\_chat\_True :

Ces caractéristiques ont une importance moindre, mais indiquent quand même que les intérêts pour faire des amis ou pour chatter peuvent jouer un rôle.

gender :

Le genre a la moindre importance parmi les caractéristiques.

**Consrtuction de la matrice de confusion**



La matrice de confusion obtenue montre les performances du modèle optimisé. Voici comment l'interpréter :

* **Axes** :
  + L'axe vertical représente les labels réels (True labels).
  + L'axe horizontal représente les labels prédits (Predicted labels).
* **Éléments de la Matrice** :
  + **True Negatives (TN)** : 387
    - Nombre de profils non populaires correctement classés comme non populaires.
  + **False Positives (FP)** : 10
    - Nombre de profils non populaires incorrectement classés comme populaires.
  + **False Negatives (FN)** : 21
    - Nombre de profils populaires incorrectement classés comme non populaires.
  + **True Positives (TP)** : 125
    - Nombre de profils populaires correctement classés comme populaires.

**Calcul des Métriques à partir de la Matrice de Confusion**

