**Projet IA :** Détection Automatique de Genre pour la Carte d'Identité Nationale du Togo

**Introduction :**

L'objectif de ce projet est de développer un modèle d'apprentissage automatique capable de prédire automatiquement le genre (masculin/féminin) d'une personne à partir de son prénom, en utilisant un **dataset** contenant des noms et prénoms associés à leur genre. Ce rapport résume les étapes clés du projet, les résultats obtenus et les visualisations pertinentes.

**Mission :**

Développement d’un système intelligent qui détecte automatiquement le genre à partir des noms/prénoms pour :

- Éliminer les erreurs humaines dans la base nationale

- Optimiser le traitement des 500K demandes annuelles

**Étapes de l'Exercice**

* **Étape 1 : Préparation des Données**

***Chargement du dataset (noms\_prenoms.csv) contenant :***

**full\_name** : une chaîne de caractères au format "Nom Prénom".

**gender** : 0 (Femme) ou 1 (Homme).

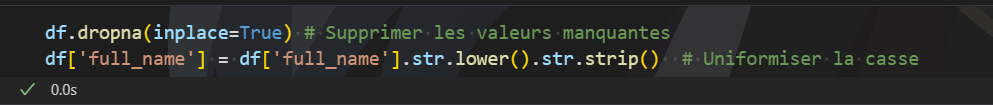
Dataset utilisé : **dataset\_names.csv** contenant les colonnes **lastname**, **firstname**, et **gender**.



**Nettoyage des données :**

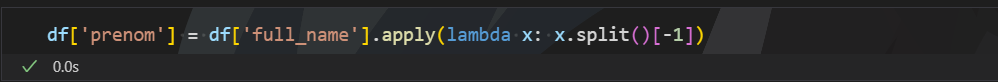
Vérification des valeurs manquantes.

Uniformisation de la casse (majuscules/minuscules).



**Feature Engineering :**

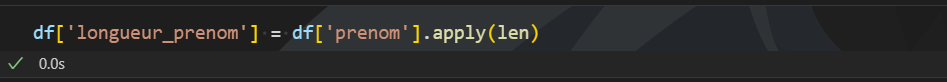
Extraction du prénom (dernier mot de **full\_name**).



**Création des features manuelles telles que :**

Les caractéristiques suivantes ont été créées pour améliorer la prédiction :

Longueur du prénom : Nombre de caractères.



Terminaisons fréquentes (ex : '**a**', '**e**' pour féminin, '**ou**', '**d**’pour masculin).

Une image contenant capture d’écran, texte, Logiciel multimédia

Le contenu généré par l’IA peut être incorrect.

Préfixes courants (ex : "Mohamed : Masculin", "Aïssa : Féminin ").

Vectorisation des prénoms : Utilisation de ***CountVectorizer*** pour les n-grammes de caractères (bigrammes et trigrammes).

Une image contenant texte, capture d’écran, Logiciel multimédia, logiciel

Le contenu généré par l’IA peut être incorrect.

* **Étape 2 : Séparation des Données**

***Split Train/Test*** ***:***

Train (70%) pour l'entraînement.

Test (30%) pour l'évaluation.

Sortie du code montrant les tailles des datasets (**X\_train.shape, X\_test.shape**).

Une image contenant texte, capture d’écran, Logiciel multimédia, logiciel

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police, noir

Le contenu généré par l’IA peut être incorrect.

Taille des jeux de données :

**Train : 7 000 échantillons.**

**Test : 3 000 échantillons.**

* **Étape 3 : Entraînement du Modèle (Ensemble Learning)**

***Trois modèles d'ensemble Learning ont été utilisés :***

* **Random Forest**
* **XGBoost**
* **Gradient Boosting** (optionnel)

Une image contenant texte, capture d’écran, Logiciel multimédia

Le contenu généré par l’IA peut être incorrect.

Entraînement des modèles sur le jeu d'entraînement.

Sortie de l'entraînement

Une image contenant texte, Police, nombre, logiciel

Le contenu généré par l’IA peut être incorrect.

* **Étape 4 : Évaluation des Performances**

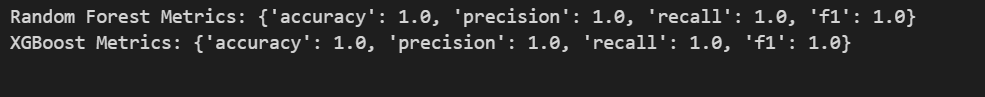
***Calcul des métriques suivantes sur le jeu de test***

Les modèles ont été évalués sur :

* Accuracy : Taux de prédictions correctes.
* Precision : Proportion de prédictions positives correctes.
* Recall : Proportion de vrais positifs détectés.
* F1-Score : Moyenne harmonique de précision et rappel.

Une image contenant texte, capture d’écran, Logiciel multimédia, logiciel

Le contenu généré par l’IA peut être incorrect.



Résultats :

Sortie des métriques (**rf\_metrics, xgb\_metrics**).

**Random Forest** et **XGBoost** ont atteint une accuracy de 100% sur le jeu de test.

**Comparation des performances des différents modèles.**

Les deux modèles principaux montrent des performances similaires, mais **XGBoost** est souvent plus efficace pour les problèmes complexes.

* **Étape 5 : Optimisation (Bonus)**

***Hyperparamétrage avec GridSearchCV pour l’optimisation des paramètres.***

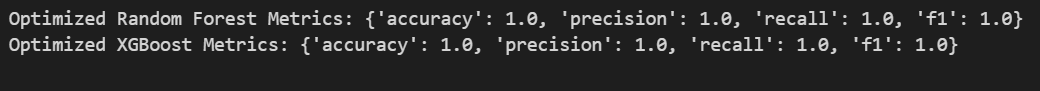
**Random Forest** : Optimisation de n\_estimators, max\_depth, min\_samples\_split.

**XGBoost** : Optimisation de n\_estimators, max\_depth, learning\_rate.

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Le contenu généré par l’IA peut être incorrect.

Meilleurs paramètres trouvés par ***GridSearchCV.***



**Feature Importance :**

Top 10 des features pour **Random Forest** :

termine\_par\_e (le plus influent).

termine\_par\_a, ia, ne, longueur\_prenom, etc.

Une image contenant texte, capture d’écran, Logiciel multimédia, Police

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, noir, Police

Le contenu généré par l’IA peut être incorrect.

Graphique des features les plus importantes

**Visualisations Clés :**

Diagramme en barres montrant l'importance des caractéristiques.

Explication : Les terminaisons de prénoms sont très discriminantes.

Pour montrer les vrais/faux positifs/négatifs.

Une image contenant texte, capture d’écran

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

**Analyse pour Random Forest :**

Les features les plus importantes sont termine\_par\_e (la plus influente), suivie par termine\_par\_a et ia. On observe que les terminaisons de prénoms et les préfixes semblent particulièrement discriminants.

Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

**Analyse pour XGBoost:**

Les features les plus importantes sont termine\_par\_a (la plus influente), suivie par termine\_par\_e et ga. On observe que les terminaisons de prénoms et les préfixes semblent particulièrement discriminants.

**Comparaison des performances des modèles.**

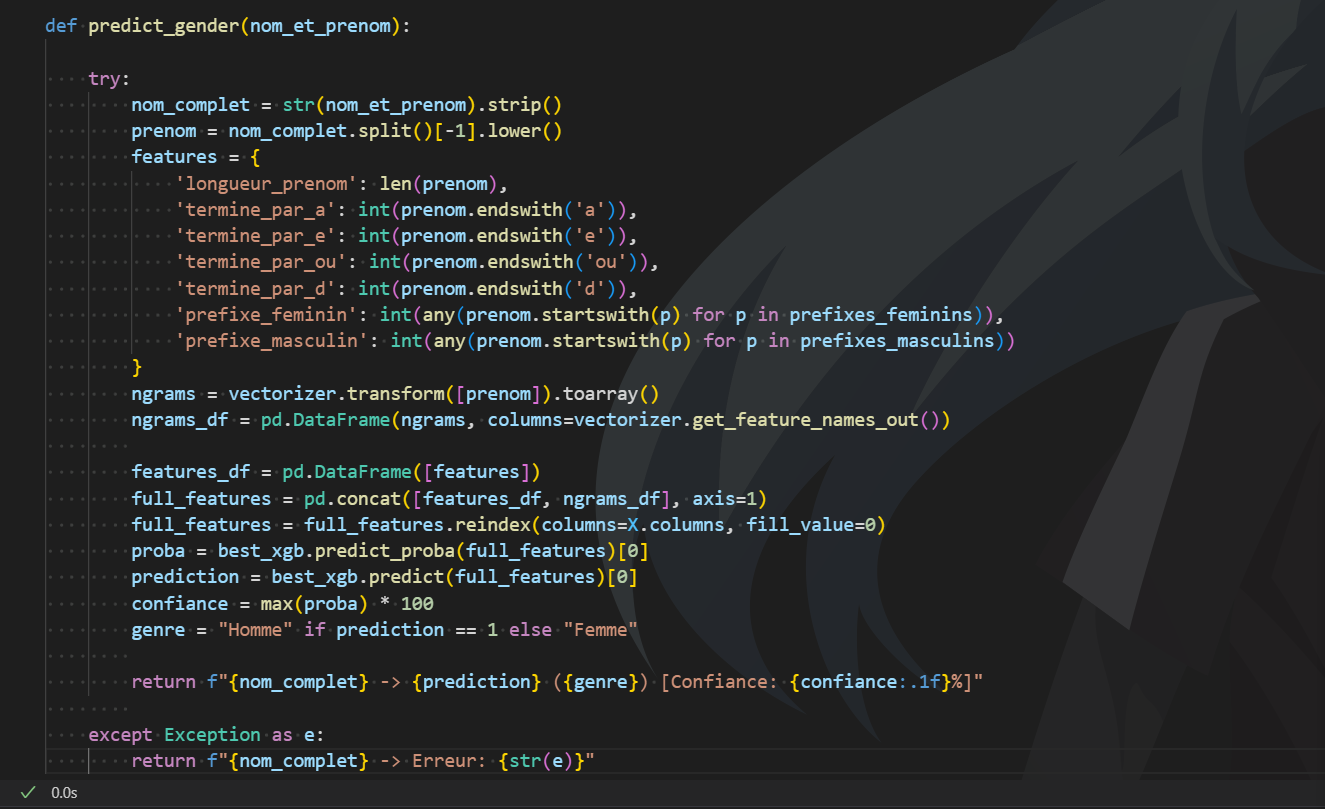
Les deux modèles s'accordent généralement sur l'importance des motifs de caractères (n-grams) mais diffèrent sur l'importance relative des features manuelles (comme la longueur du prénom). Cette analyse peut guider la sélection de features pour de futurs modèles.

* **Étape 6 : Prédiction sur de Nouvelles Données**

**Fonction *predict\_gender(nom\_et\_prenom)* qui retourne :**

0 si le modèle prédit Femme.

1 si le modèle prédit Homme.



Exemple de cas d’utilisation de l’algorithme

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police, conception

Le contenu généré par l’IA peut être incorrect.