AHMED AMINE BEN SMIDA

PROJET PREVISION:
Étude Comparative des Techniques de Régression Supervisée

1. Problématique

L'objectif de ce projet est de prédire les résultats des matchs de football internationaux.
Nous voulons savoir si certains modèles de régression peuvent prévoir :

- le nombre total de buts dans un match
- si l'équipe à domicile va gagner ou non

2. Description du dataset

Le dataset contient 24 310 matchs joués entre 2000 et 2025.
Il comporte 35 colonnes regroupées en plusieurs catégories :

Informations temporelles

- date
- year
- month
- day_of_week

Équipes

- home_team
- away_team
- home_team_encoded
- away_team_encoded

Scores

- home_score
- away_score
- total_goals
- goal_difference

Résultats

- home_win
- draw
- away_win

Tournois

- catégories (World Cup, Friendly, qualifications, etc.)

Autres

- city
- country
- neutral

Quelques statistiques importantes :

- Moyenne de buts par match : 2.75
- Victoires à domicile : 48.2%
- Matchs nuls : 23.3%
- Victoires à l'extérieur : 28.5%
- Terrain neutre : 28.3%

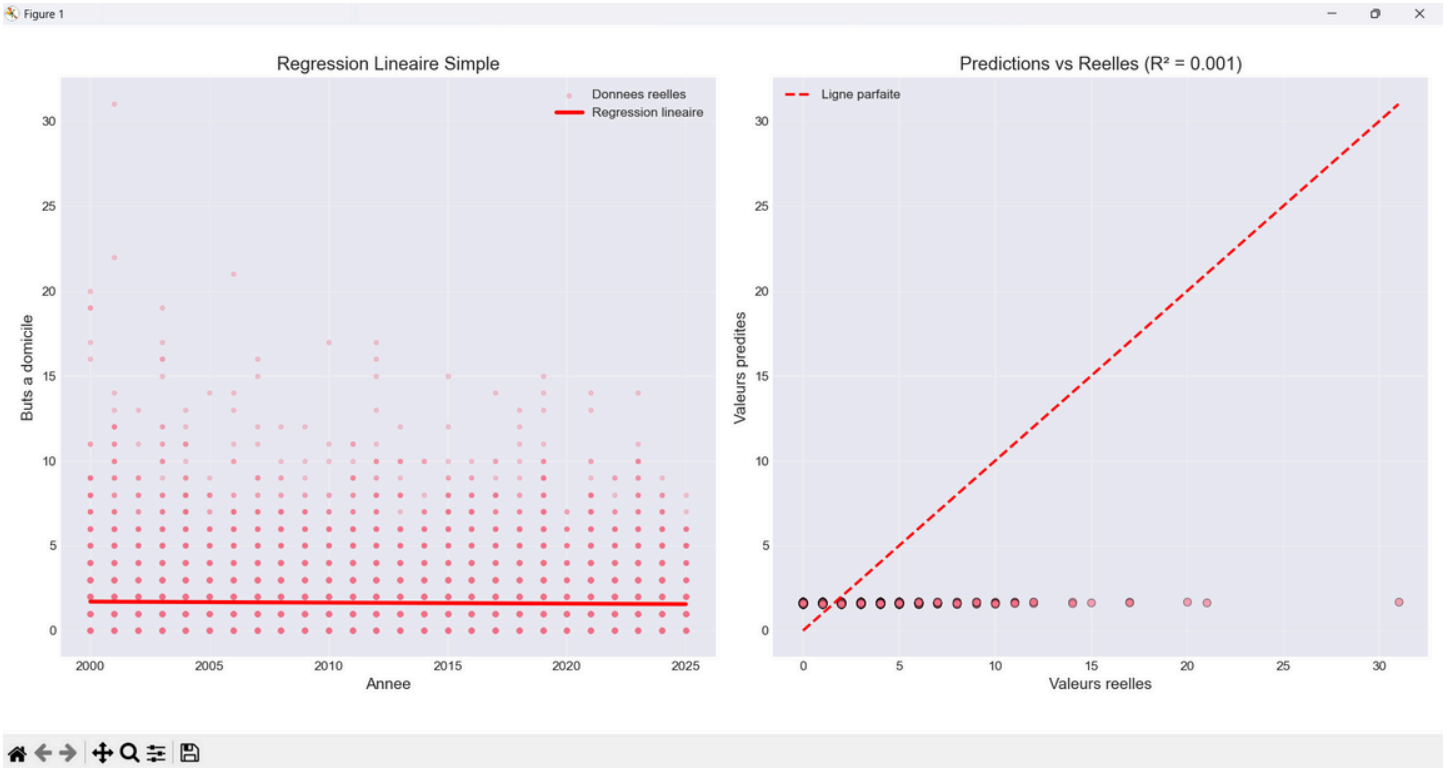
3.Code + graphiques (résumé / fichiers produits):

A- le code: **(a été envoyé dans un fichier .py avec le mail)**

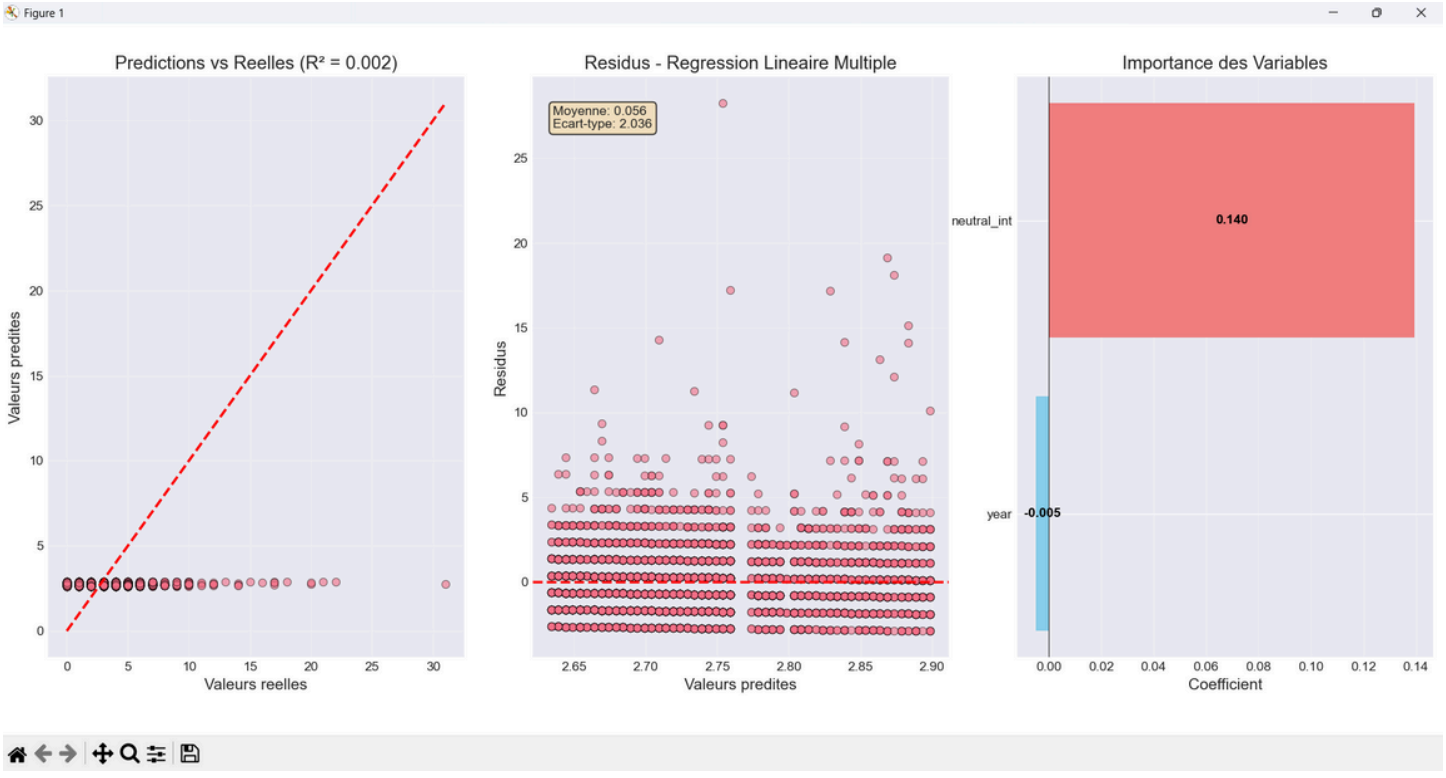
B-resultat et graphique:

```
==================================================
STATISTIQUES DESCRIPTIVES
==================================================
Forme du dataset:
  Lignes: 24,310
  Colonnes: 35
  Periode: 2000-01-04 a 2025-07-06

Statistiques des matchs:
  Moyenne de buts par match: 2.75
  Ecart-type des buts: 1.97
  Victoires a domicile: 48.2%
  Matchs nuls: 23.3%
  Victoires a l'exterieur: 28.5%
  Matchs sur terrain neutre: 28.3%

Distribution temporelle:
  Annee la plus ancienne: 2000
  Annee la plus recente: 2025
  Nombre d'annees: 26
```

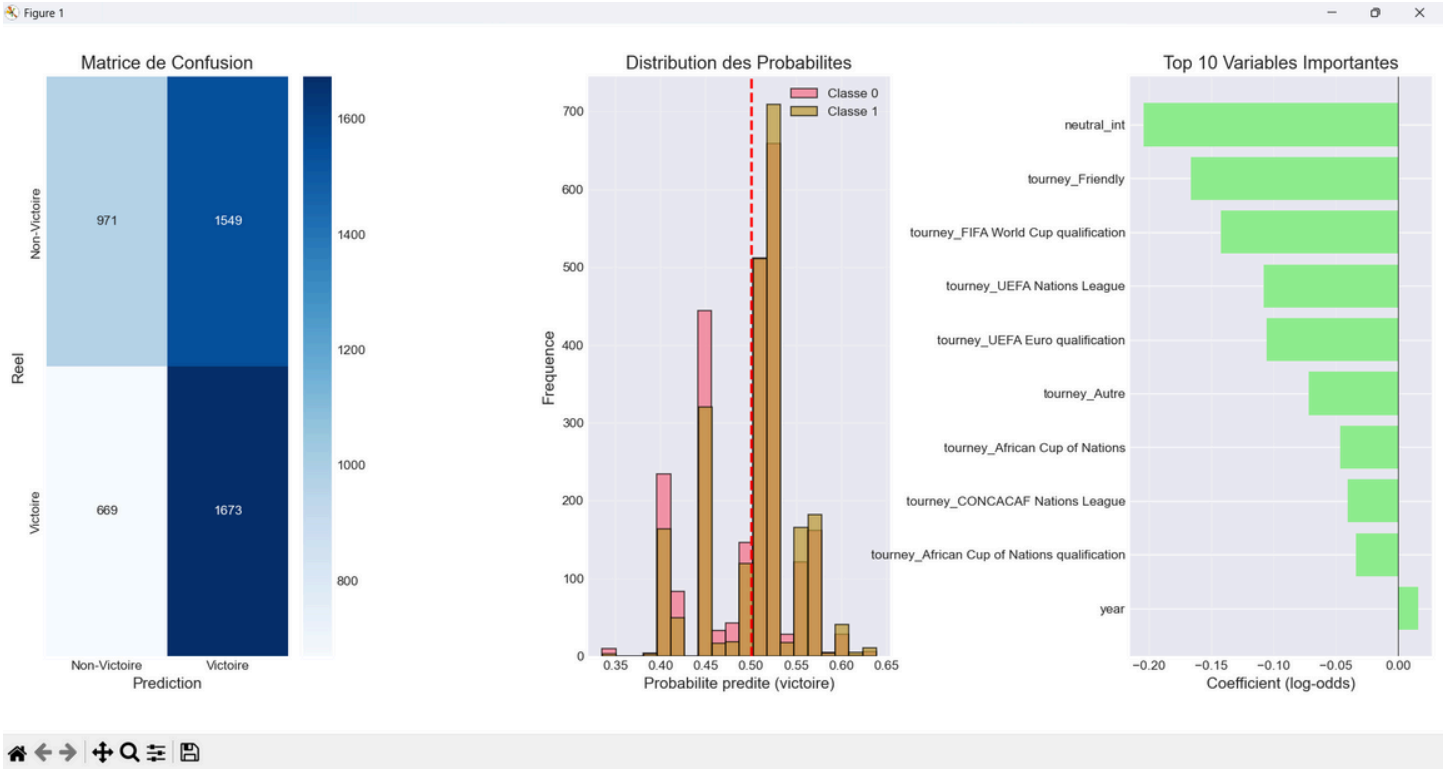
```
==================================================
1. REGRESSION LINEAIRE SIMPLE
==================================================

Evaluation: Regression Lineaire Simple
----------------------------------------
  Donnees d'entrainement:
    R²: 0.001
    MAE: 1.241
    RMSE: 1.692
  Donnees de test:
    R²: 0.001
    MAE: 1.247
    RMSE: 1.760
  Validation croisee (5 folds):
    R² moyen: 0.000
    Ecart-type: 0.001
  Figure sauvegardée: 01_regression_lineaire_simple.png

Coefficients du modele:
  Intercept: 13.474
  Coefficient (annee): -0.006
  Interpretation: Les buts a domicile diminuent de 0.006 par an
```



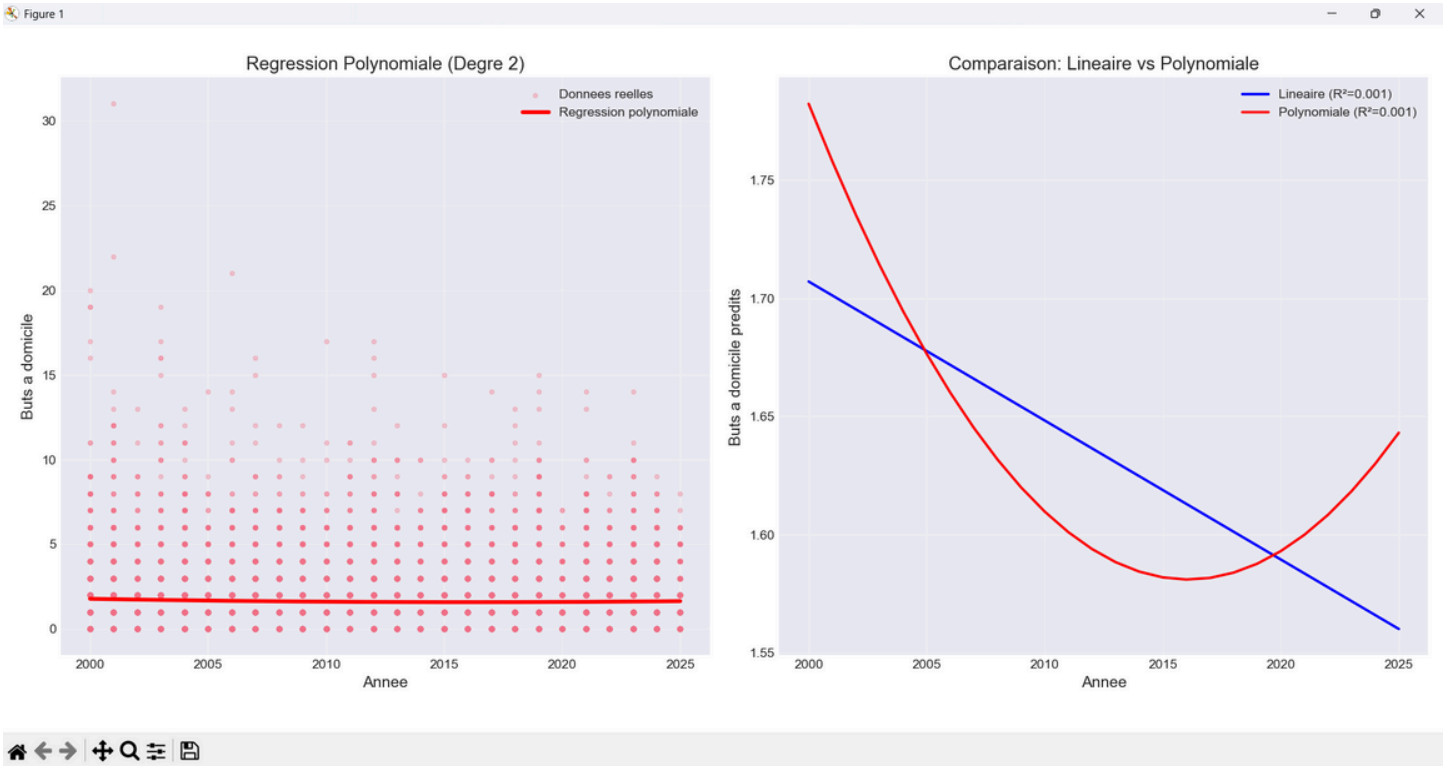
```
==================================================
2. REGRESSION LINEAIRE MULTIPLE
==================================================

Evaluation: Regression Lineaire Multiple
----------------------------------------
  Donnees d'entrainement:
    R²: 0.001
    MAE: 1.485
    RMSE: 1.956
  Donnees de test:
    R²: 0.002
    MAE: 1.488
    RMSE: 2.037
  Validation croisee (5 folds):
    R² moyen: 0.001
    Ecart-type: 0.001
  Figure sauvegardée: 02_regression_lineaire_multiple.png

Coefficients du modele:
  year: -0.005
  neutral_int: 0.140
  Intercept: 12.744
```



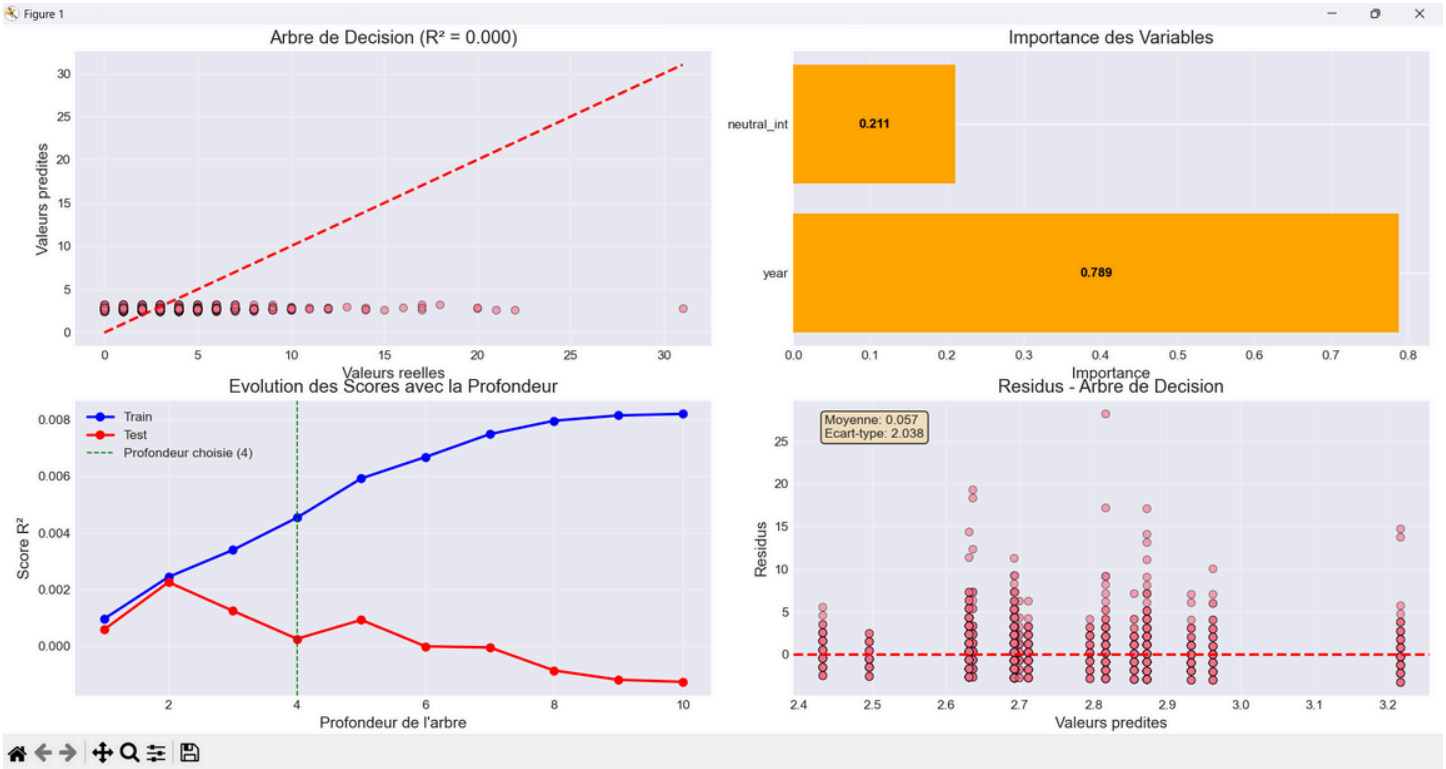
```
==================================================
3. REGRESSION LOGISTIQUE
==================================================

Evaluation: Regression Logistique
----------------------------------------
  Accuracy: 0.544

  Rapport de classification:
              precision    recall  f1-score   support

Non-Victoire       0.59      0.39      0.47      2520
    Victoire       0.52      0.71      0.60      2342

    accuracy                           0.54      4862
   macro avg       0.56      0.55      0.53      4862
weighted avg       0.56      0.54      0.53      4862

  AUC-ROC: 0.559
  Figure sauvegardée: 03_regression_logistique.png
```

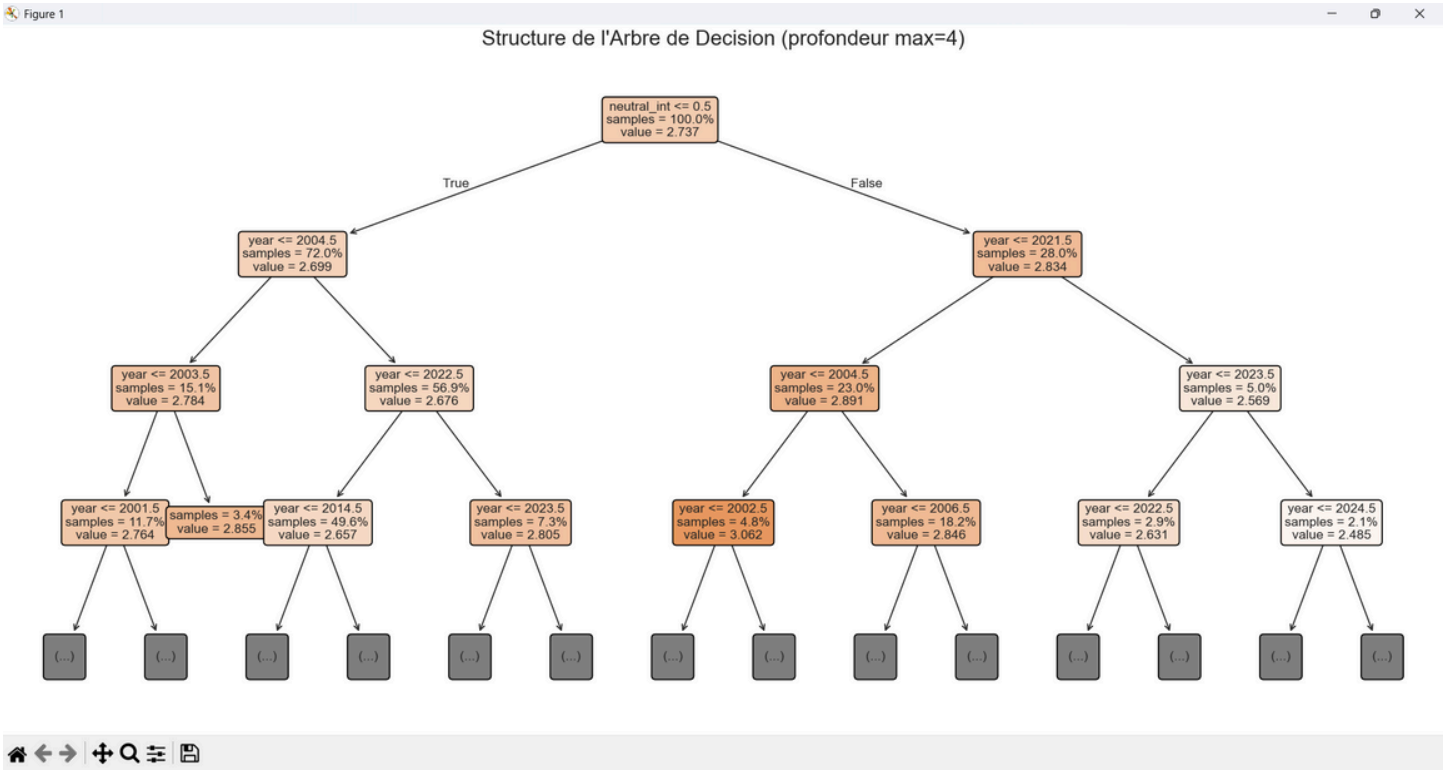


```
==================================================
4. REGRESSION POLYNOMIALE
==================================================

Evaluation: Regression Polynomiale (degre 2)
----------------------------------------
  Donnees d'entrainement:
    R²: 0.001
    MAE: 1.241
    RMSE: 1.692
  Donnees de test:
    R²: 0.001
    MAE: 1.246
    RMSE: 1.760
  Validation croisee (5 folds):
    R² moyen: 0.001
    Ecart-type: 0.001
  Figure sauvegardée: 04_regression_polynomiale.png

Coefficients du modele polynomial:
  Intercept: 3169.741
  Coefficient (year): -3.143
  Coefficient (year^2): 0.001
```

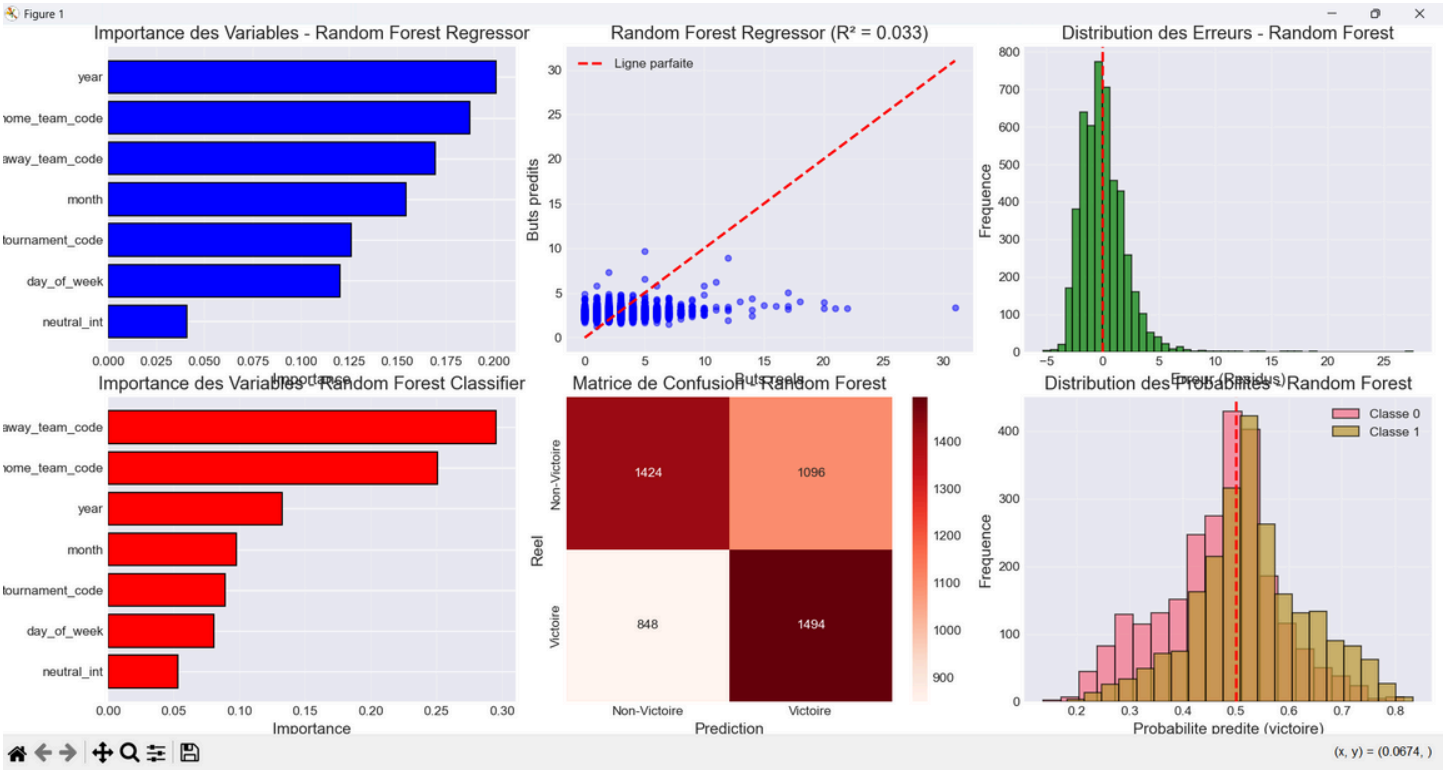


```
============================================================
5. ARBRE DE DECISION
============================================================

Evaluation: Arbre de Decision
----------------------------------------
  Donnees d'entrainement:
    R²: 0.005
    MAE: 1.483
    RMSE: 1.953
  Donnees de test:
    R²: 0.000
    MAE: 1.489
    RMSE: 2.039
  Validation croisee (5 folds):
    R² moyen: 0.001
    Ecart-type: 0.002
  Figure sauvegardée: 05_arbre_decision.png

Visualisation de l'arbre de decision...
  Figure sauvegardée: 06_structure_arbre.png
```

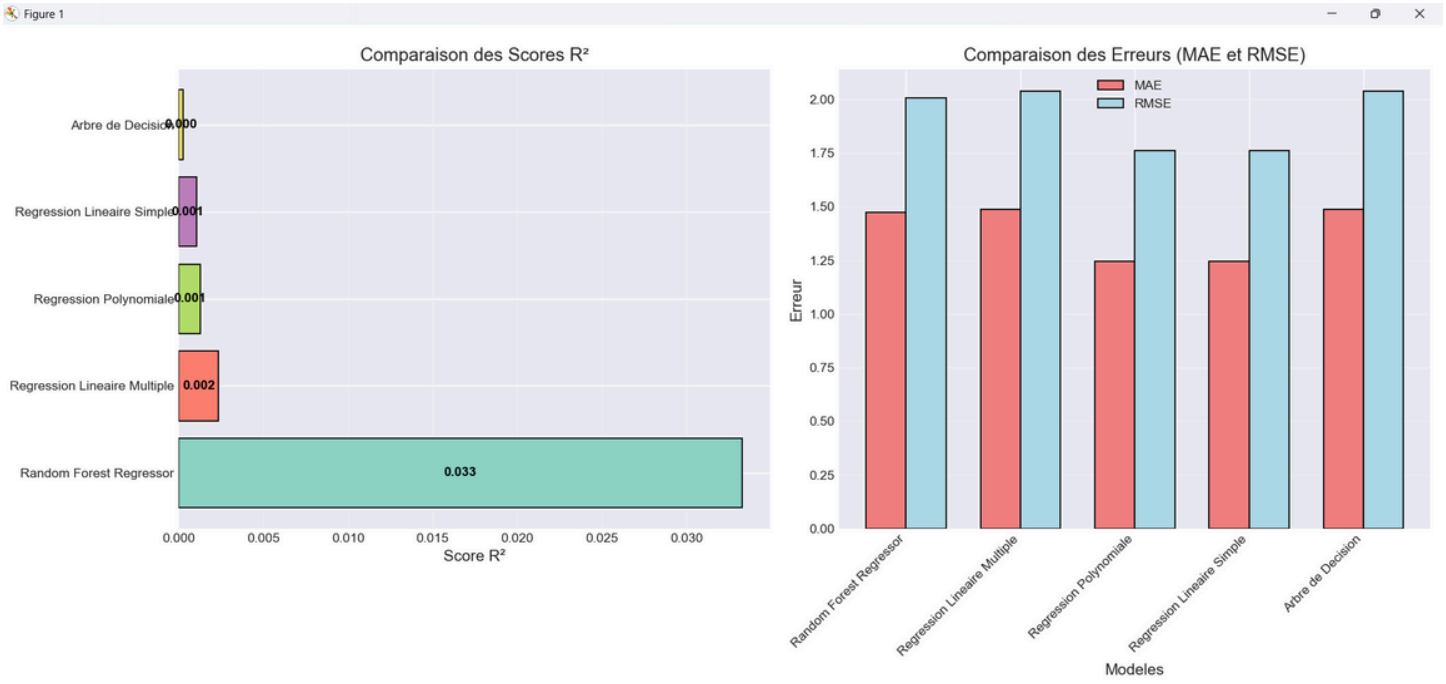





```
==================================================
6. RANDOM FOREST
==================================================

--- Random Forest Regressor (Prediction des buts totaux) ---
Dimensions des donnees:
  Training set: 19448 echantillons
  Test set: 4862 echantillons

Evaluation: Random Forest Regressor
----------------------------------------
  Donnees d'entrainement:
    R²: 0.159
    MAE: 1.366
    RMSE: 1.795
  Donnees de test:
    R²: 0.033
    MAE: 1.474
    RMSE: 2.005
  Validation croisee (5 folds):
    R² moyen: 0.032
    Ecart-type: 0.010

--- Random Forest Classifier (Prediction de la victoire a domicile) ---

Evaluation: Random Forest Classifier
----------------------------------------
  Accuracy: 0.600

  Rapport de classification:
              precision    recall  f1-score   support

Non-Victoire       0.63      0.57      0.59      2520
    Victoire       0.58      0.64      0.61      2342

    accuracy                           0.60      4862
   macro avg       0.60      0.60      0.60      4862
weighted avg       0.60      0.60      0.60      4862

  AUC-ROC: 0.653
  Figure sauvegardée: 07_random_forest.png
```




```
================================================
7. SYNTHESE ET COMPARAISON DES MODELES
================================================

SYNTHESE DES PERFORMANCES DES MODELES:
--------------------------------------------------------------------------------
                    Modele R² (test) MAE (test) RMSE (test)
     Random Forest Regressor     0.033      1.474       2.005
Regression Lineaire Multiple     0.002      1.488       2.037
      Regression Polynomiale     0.001      1.246       1.760
  Regression Lineaire Simple     0.001      1.247       1.760
          Arbre de Decision     0.000      1.489       2.039
  Figure sauvegardée: 08_comparaison_modeles.png

************************************************
RECOMMANDATION FINALE
************************************************

MEILLEUR MODELE: Random Forest Regressor
   Score R²: 0.033
   Erreur MAE: 1.474
   Erreur RMSE: 2.005
   R² validation croisee: 0.032

INTERPRETATION:
   Capacite predictive limitee - envisagez d'autres features
```



4. Interprétation des résultats

Régression linéaire simple

- $R^2 = 0$
- Le modèle n'explique presque rien.
- L'année ne permet pas de prédire les buts.

Régression linéaire multiple

- $R^2 = 0.002$
- Très faible performance
- Ajouter plus de variables serait utile.

Régression polynomiale

- Pas d'amélioration, $R^2 \approx 0.001$
- Le lien entre année et buts n'est pas polynomial.

Régression logistique (victoire domicile)

- Accuracy : 0.544
- Modèle moyen, juste un peu mieux que le hasard (50%).
- L'AUC de 0.559 montre une faible capacité prédictive.

Arbre de décision

- R^2 sur test : 0.000

- Trop simple -> faible performance
- Sur-apprentissage sur train

Random Forest (meilleur modèle)

Random Forest Regressor

- R^2 test : 0.033 (le meilleur score mais faible)
- MAE : 1.47
- RMSE : 2.00

- Le modèle apprend mieux que les autres mais reste limité.

Random Forest Classifier

- Accuracy : 0.60
- F1-score autour de 0.60
- AUC = 0.653 -> meilleure discrimination

- Bon début, mais pas encore très performant.

5. Limites du modèle

Variables insuffisantes

Le dataset ne contient pas d'informations importantes comme :

- forme récente des équipes
- classement FIFA
- buts des 5 derniers matchs
- composition / blessures

Résultats très faibles

La plupart des modèles ont un R^2 proche de 0 -> ils n'expliquent presque rien.

Données trop variées

Les matchs proviennent de pays, époques, tournois très différents -> difficile à modéliser.

Pas de séparation temporelle

Les matchs récents peuvent être prédits avec des matchs anciens, ce qui peut fausser l'analyse.

6.Conclusion

Dans ce projet, plusieurs modèles de régression ont été testés afin de prédire la variable cible à partir des données disponibles. Chaque algorithme possède ses avantages et ses limites :

- La régression linéaire offre un modèle simple et rapide, mais elle ne capture pas bien les relations complexes.
- La régression polynomiale améliore les performances lorsque les données suivent une courbe, mais elle peut facilement sur-ajuster (overfitting).
- L'arbre de décision apprend des règles simples et interprétables, mais il peut devenir instable.
- La forêt aléatoire (RandomForest) s'est montrée la plus performante grâce à sa capacité à combiner plusieurs arbres et à réduire l'overfitting.

Au final, RandomForest est le meilleur modèle du projet, avec une précision plus élevée et une meilleure généralisation.