

TENN'ACE



PROJET DATA

PRÉDICTION DES ACES DANS UN MATCH DE TENNIS

Mathias Buée, Paul Consalès, Marie
Simon, Elliott Ollivier, Timothée
Henriot

PRÉSENTATION DE L'ENTREPRISE



QUI SOMMES NOUS

Une entreprise spécialisée dans **l'analyse prédictive pour le tennis**, utilisant l'apprentissage machine pour **anticiper le nombre d'aces**

NOTRE MISSION

Offrir des prédictions précises sur les aces pour aider **les bookmakers, parieurs professionnels, journalistes et analystes** à prendre de meilleures décisions.

INTRODUCTION AU PROJET



- **Contexte** : Le service est un des coups les plus déterminants au tennis. Un ace, non seulement donne un point directement mais met aussi une pression psychologique sur l'adversaire.
- **Objectif** : Notre projet vise à utiliser des modèles de machine learning pour prédire la probabilité d'aces durant les matchs de tennis, afin d'évaluer les cotes sportives et identifier les paris de valeur contre les bookmakers

INTRODUCTION

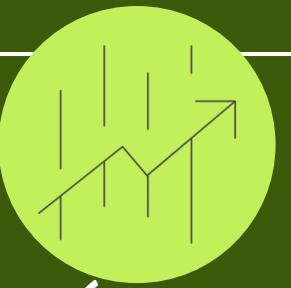




PROBLÉMATIQUE

Comment un modèle d'apprentissage peut-il améliorer la précision des prédictions Over/Under sur le nombre d'aces dans un match de tennis afin d'optimiser les décisions des parieurs et des opérateurs de paris sportifs ?

PROBLÉMATIQUE BUSINESS



MARCHÉ DES PARIS
SPORTIFS



BOOKMAKER

BETCLIC, WINAMAX...

Ajustement des cotes de manière plus fine, renforcement de la compétitivité sur le marché des paris sportifs.



PARIEURS

PARTICULIER, CURIEUX...

Maximisation des gains grâce à l'exploitation d'éventuelles imprécisions dans les cotes des bookmakers



ANALYSTES

JOURNALISTES, COACHES...

Analyse de la performance des joueurs

STRATÉGIE BUSINESS

MARCHÉ DES PARIS SPORTIFS



Comment évaluer avec précision les probabilités d'aces dans les paris Over/Under au tennis ?



Approches traditionnelles

S'appuyer sur l'expérience et les statistiques générales



Modèles d'apprentissage automatique

Analyser les données récentes des joueurs pour des prévisions précises



PROBABILITÉS ET CÔTES BOOKMAKER

$$P_i = 1 / c$$

PROBABILITE IMPLICITE
ÉMISE PAR LE BOOKMAKER

VALEUR=
 $(\text{PROBABILITÉ ESTIMÉE}) \times (\text{COTE}) - 1$

car $P_{\text{estimée}} > P_i$

VALEUR > 0

STATISTIQUEMENT
RENTABLE

APPROCHE LONG TERME

MARCHÉ ANALYSÉ : *OVER/UNDER*

6 BOOKMAKER ANJ EN FRANCE

MÉTHODOLOGIE DE L'ÉTUDE

CHARGEMENT ET EXPLORATION DES DONNÉES

Importer les données (lire le fichier `atp_matches_2023.csv`) et comprendre leur structure avant de commencer la modélisation.

DÉTECTION ANOMALIES, PRÉTRAITEMENT DES DONNÉES

Identification des valeurs aberrantes, Nettoyage des valeurs manquantes, encodage des variables catégorielles (surface, joueurs), standardisation

SÉLECTION DES VARIABLES ET SÉPARATION DES DONNÉES

Identification des features les plus importantes. Division 80/20.

MODÉLISATION ET ÉVALUATION

Choix du modèle principal Random Forest. Analyse de la MAE (Mean Absolute Error) et la RMSE (Root Mean Squared Error) pour avoir une idée de la fiabilité.

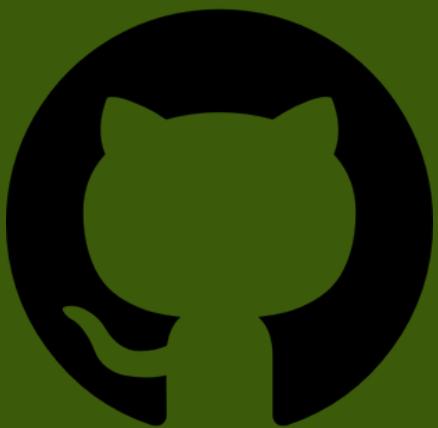


PIPELINE DE DONNÉES

Récupération live depuis GitHub :
https://github.com/JeffSackmann/tennis_atp

Actualisation quotidienne

Concaténation depuis 2010



LES DONNÉES



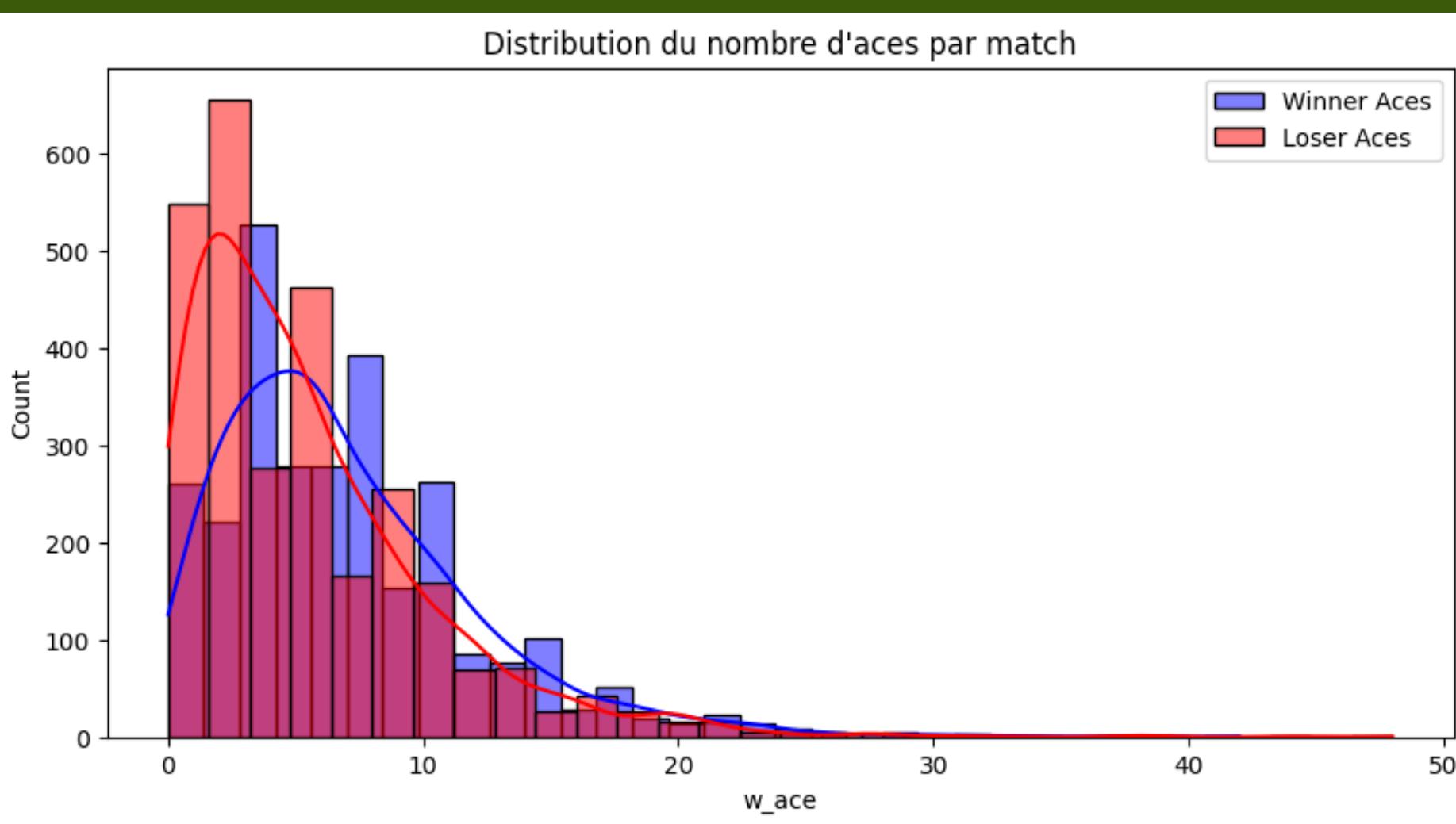
2980 MATCHS ANALYSÉS

49 VARIABLES

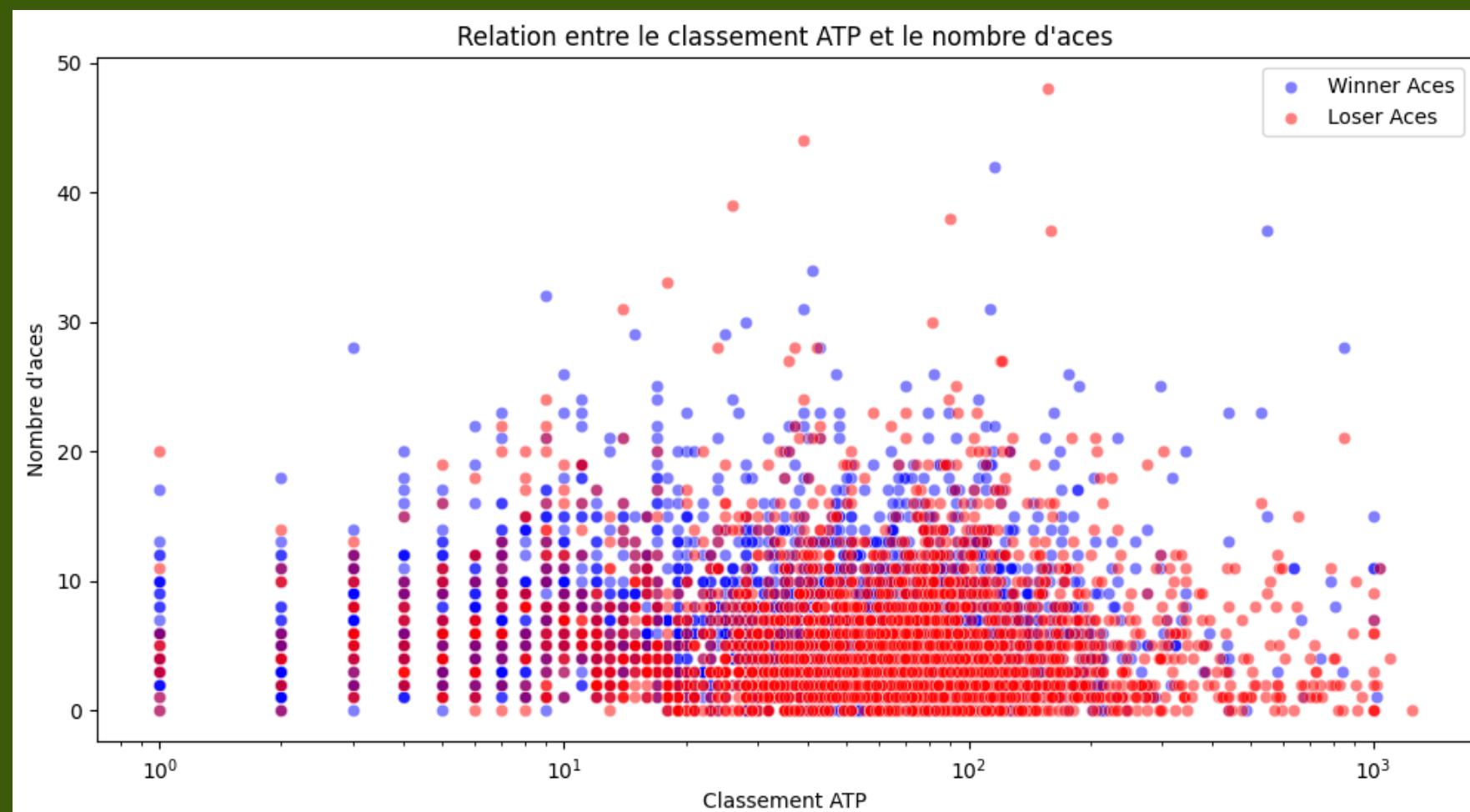
6,9 ACES/MATCH

144 TOURNOIS

EXPLORATION

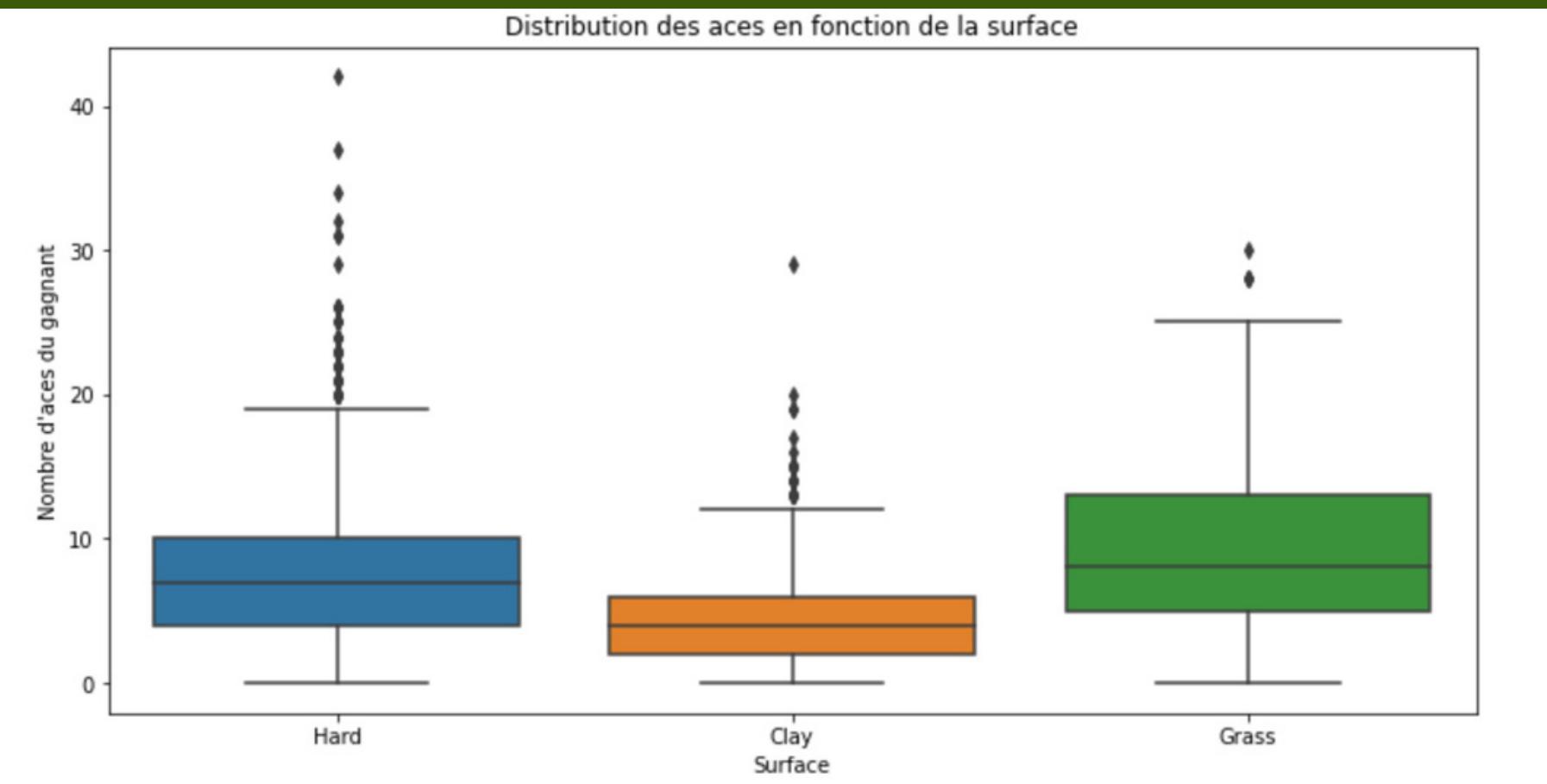


- La majorité des aces par match se situe entre **0 et 10**.
- Les gagnants réalisent généralement plus d'aces que les perdants.
- Pic de distribution à **environ 5 aces pour les gagnants**.
- Distributions asymétriques avec une longue traîne vers des nombres plus élevés d'aces.
- **Moins de 5% des joueurs réalisent plus de 20 aces par match.**
- 75% des perdants contre 60% des gagnants réalisent 5 aces ou moins, montrant une corrélation entre aces et victoires.

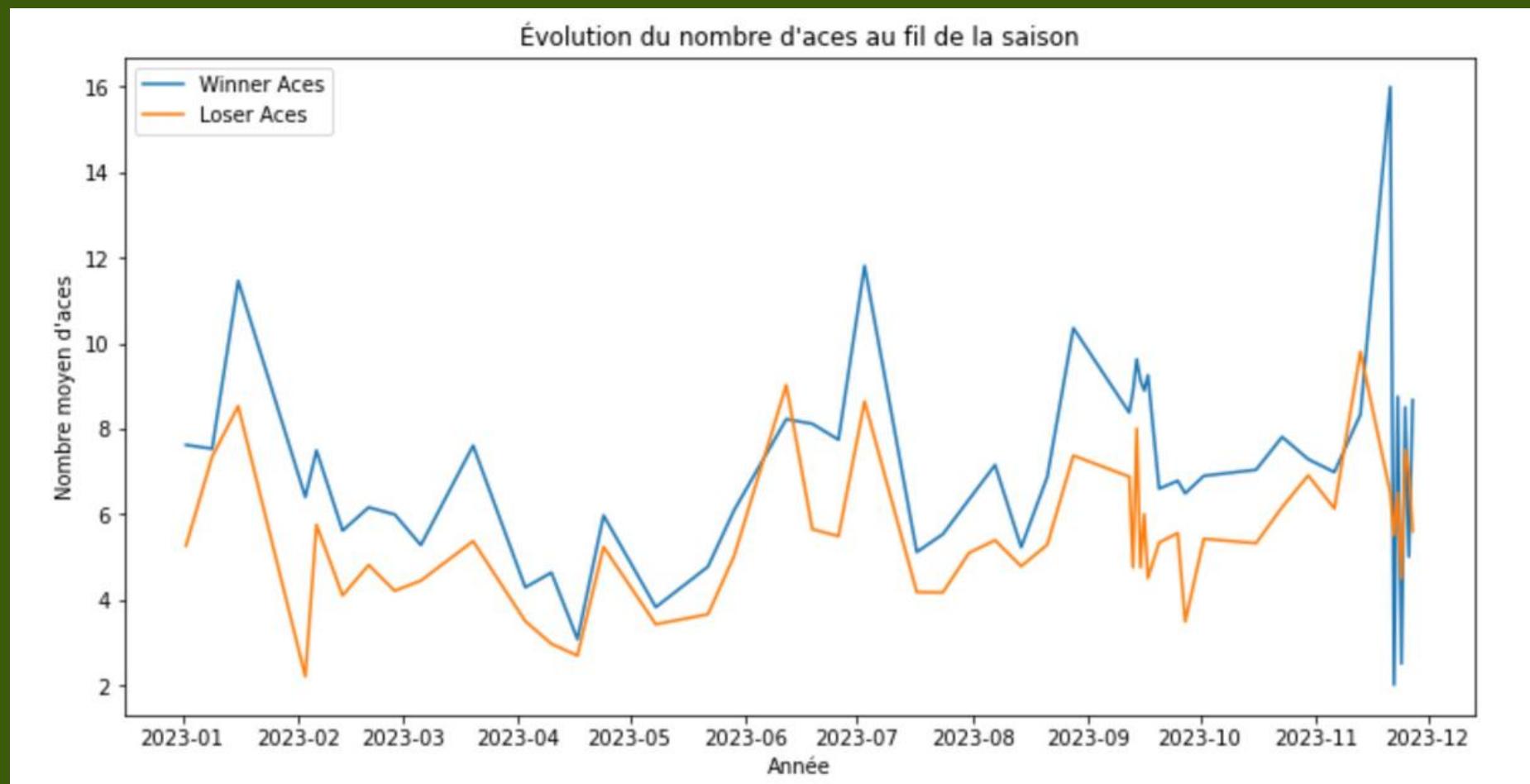


- Indépendance du Classement : **Le classement élevé ne garantit pas plus d'aces**; des facteurs autres que le classement influencent les aces.
- Variabilité des Aces : **Dispersion des aces visible à tous les niveaux de classement**, avec des performances élevées même chez des joueurs de rang inférieur.
- Distribution Variée : **Joueurs de tous niveaux**, y compris ceux classés entre 100-300, **affichent des nombres élevés d'aces**, souvent liés à des styles de jeu uniques ou des conditions favorables.

EXPLORATION

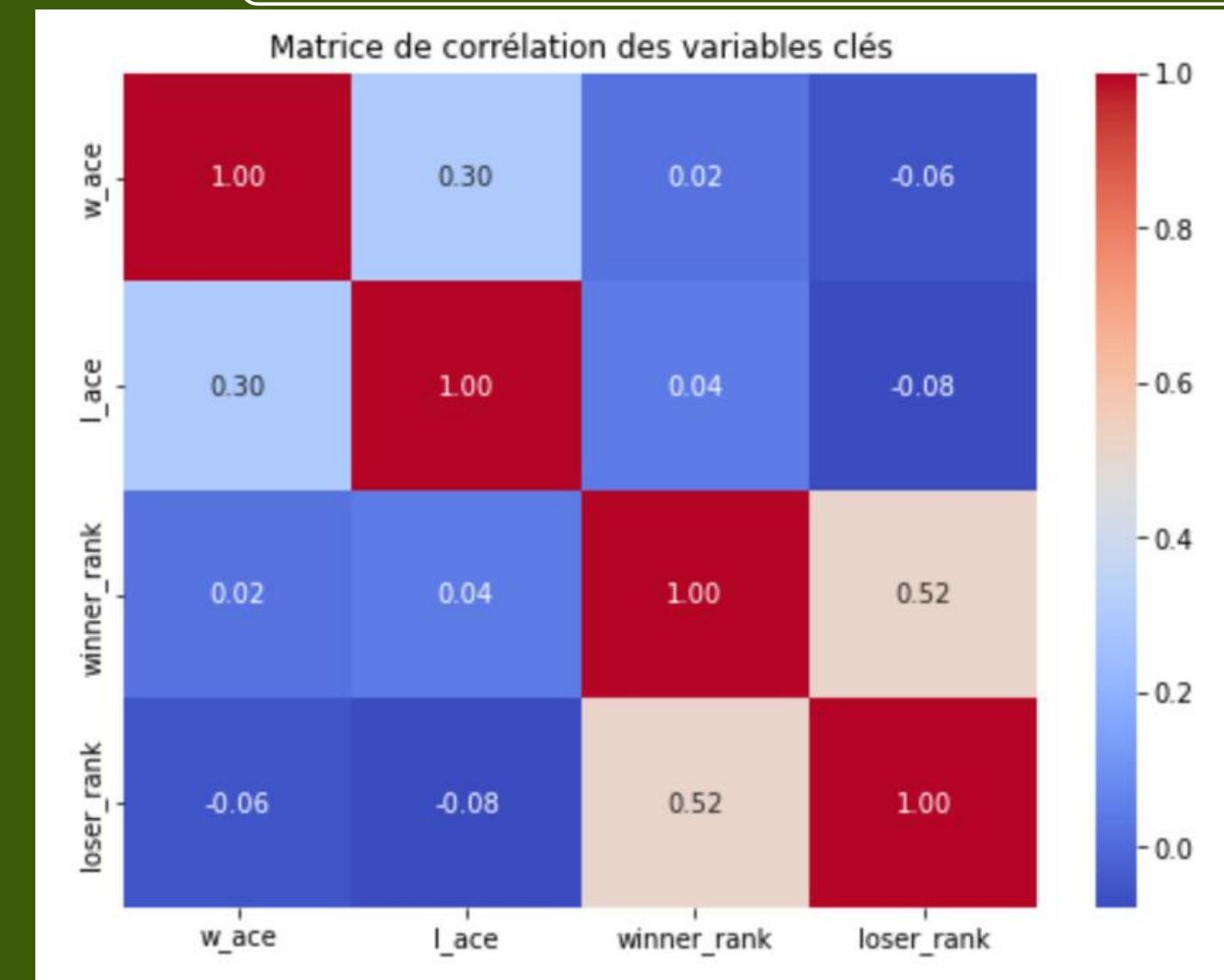
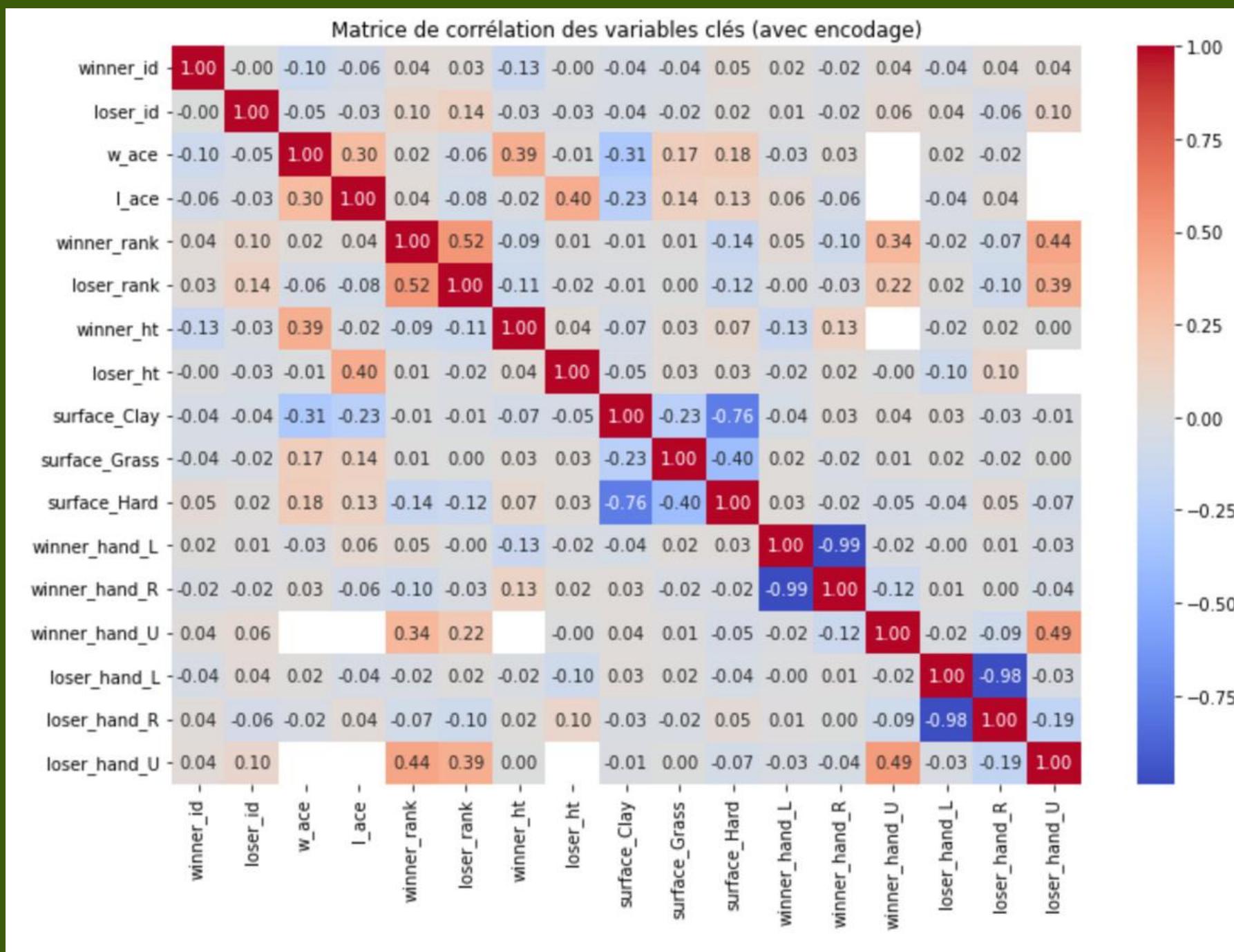


- **Hard Surface:** Médiane basse avec quelques valeurs élevées. **Moins favorable aux aces en raison de la vitesse moyenne et du rebond prévisible de la balle**, permettant aux adversaires de mieux retourner les services.
- **Clay Surface:** Variabilité élevée, reflet des défis posés par la **surface lente qui absorbe plus l'énergie de la balle**, réduisant l'efficacité des services puissants.
- **Grass Surface:** Médiane et dispersion élevées, indiquant une **préférence pour les aces**. La surface rapide et le faible rebond de la balle facilitent les services directs et difficiles à retourner.



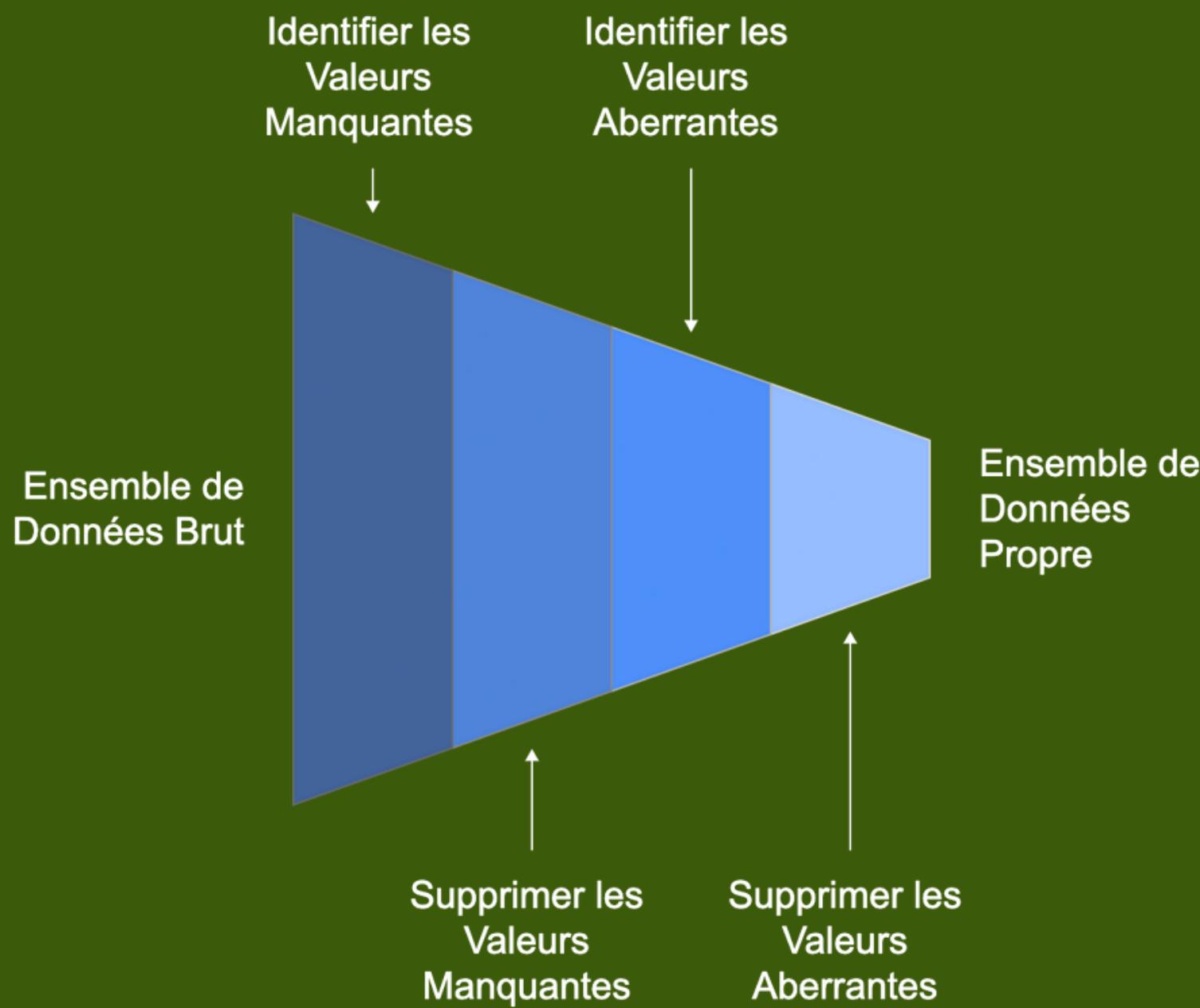
- Juin/Juillet et Septembre: **Augmentation des aces durant les Grands Chelems** (Roland-Garros en juin, US Open en septembre), liée aux formats en trois sets gagnants.
- Fin d'Année: **Hausse des aces durant les tournois indoor et la finale de l'ATP**, joués sur des surfaces rapides.
- **Wimbledon (juillet) et US Open : Surfaces rapides favorisent les services puissants**, entraînant plus d'aces.
- Les joueurs ajustent leur jeu pour maximiser les aces durant ces tournois majeurs, illustrant l'importance de la stratégie de service.

EXPLORATION



- Surface et Aces :** Forte corrélation négative entre les aces et les surfaces dures (-0.76), indiquant que **moins d'aces sont réalisés sur des surfaces dures comparées à la terre battue et au gazon.**
- Classement et Aces :** Corrélation positive modérée entre le classement du joueur et les aces réalisés par les perdants (0.52), suggérant que **les joueurs moins bien classés réalisent également moins d'aces.**
- Aces des Gagnants vs. Perdants :** Corrélation positive (0.30) entre les aces des gagnants et ceux des perdants, **montrant que les matchs avec de nombreux aces par les gagnants ont tendance à avoir également plus d'aces par les perdants.**
- Impact du Classement :** **Très faible influence du classement sur le nombre d'aces**, avec des corrélations presque négligeables (-0.06 avec les aces des gagnants et -0.08 avec les aces des perdants).

Nettoyage des données



PRÉTRAITEMENT DES DONNÉES

Transformation des données

- Création nouvelles variables (moyenne des performances...)
- Encodage variables catégorielles (surface, niveau du tournoi, format du match)
- Normalisation des variables continues (âge, taille, classement, points ATP)

Structuration pour l'apprentissage

- Attribution aléatoire des joueurs en tant que "Joueur 1" et "Joueur 2" pour éviter un biais
- Définition de la variable cible : le nombre d'ace de chaque joueur

TEST DU MODÈLE



1

Régression linéaire

Objectif : Prédire directement le nombre d'aces.

Problème : Les résultats sont peu précis en raison d'une forte dispersion des aces.

2

Ajout de l'historique des 10 derniers matchs

Objectif : Améliorer la précision en intégrant des tendances récentes.

Problème : Peu d'amélioration observée, sûrement dû à un manque de données complètes.

3

Modèle basé uniquement sur les caractéristiques des joueurs

Objectif : Se baser uniquement sur le profil des joueurs (taille, classement, style de jeu).

Problème : Les conditions de jeu (surface, météo, altitude) ont un impact trop important pour ignorer ces facteurs.

MODÈLE DE MACHINE LEARNING

COEF LASSO

Vérification des corrélations entre les variables afin de choisir les meilleurs hyperparamètres



RANDOM FOREST CLASSIFIER

Repose sur la combinaison de plusieurs arbres de décision pour améliorer la robustesse des prédictions



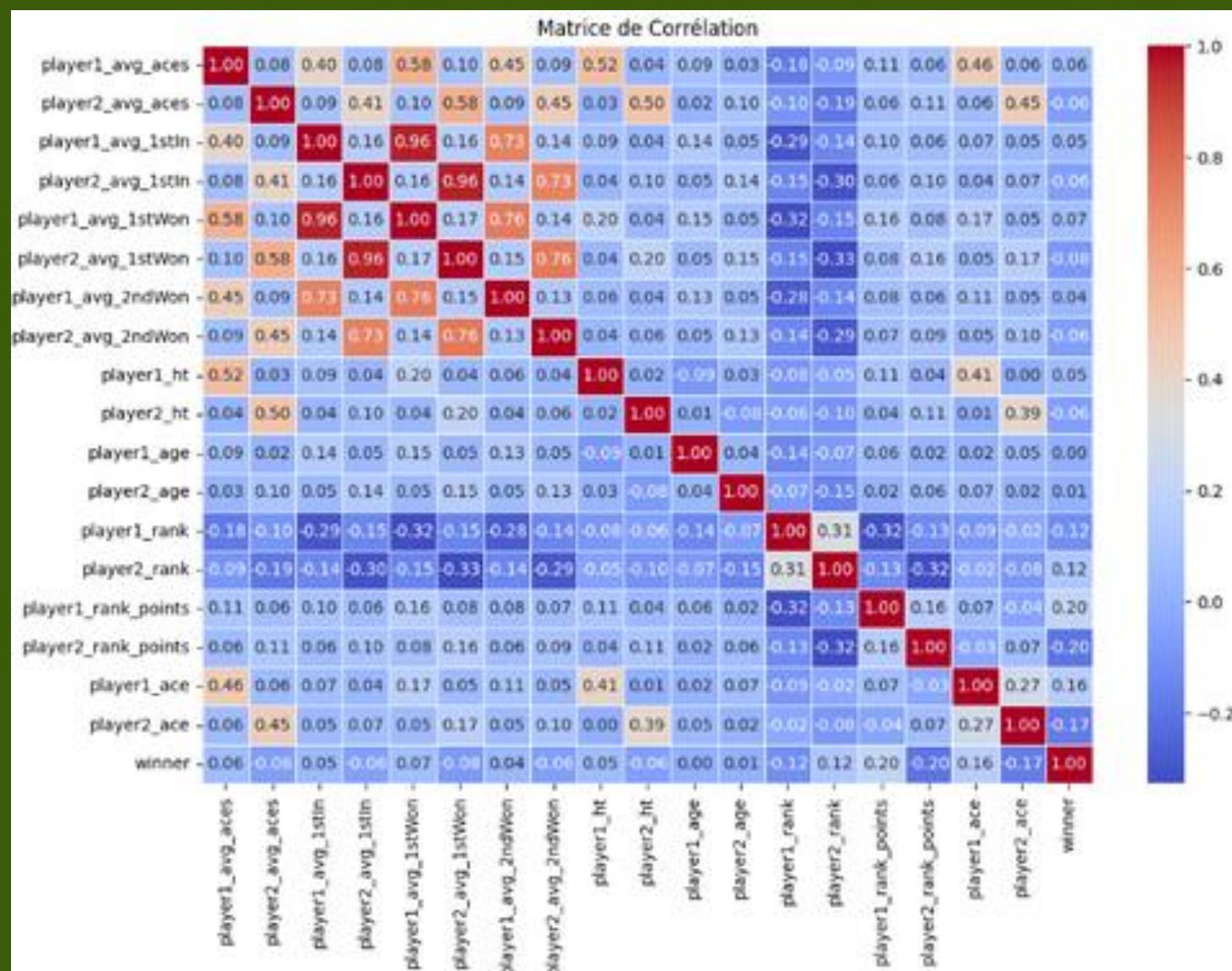
RÉSULTATS

Feature Importance : surface de jeu, classement ATP, la taille des joueurs

Amélioration possible avec des modèles plus avancés



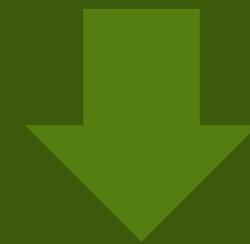
RANDOM FOREST - ENTRAINEMENT



On cherche à trouver **player1_ace** (le nombre d'ace du joueur 1 dans le match)

On **entraîne le modèle** à partir des variables ayant un **coefficent élevé** (vérifié avec Lasso)

player1 avg aces (0.37), player1 ht (0.20),
player1 1stWon, player1 1stIn, player1 rank



On trouve des résultats satisfaisants :

MAE : 0.589
RMSE : 0.818

RANDOM FOREST - EXEMPLE

```
# Création du DataFrame avec les nouvelles valeurs
new_match = pd.DataFrame({
    "player1_avg_aces": [18],
    "player1_avg_1stwon": [49],
    "player1_ht": [188],
    "player2_ace": [10],
    "player1_avg_2ndwon": [21],
    "player1_avg_1stIn": [67],
    "player1_rank": [70]
})

# Ensure new_match only contains selected features before scaling
new_match_selected = new_match[selected_features]

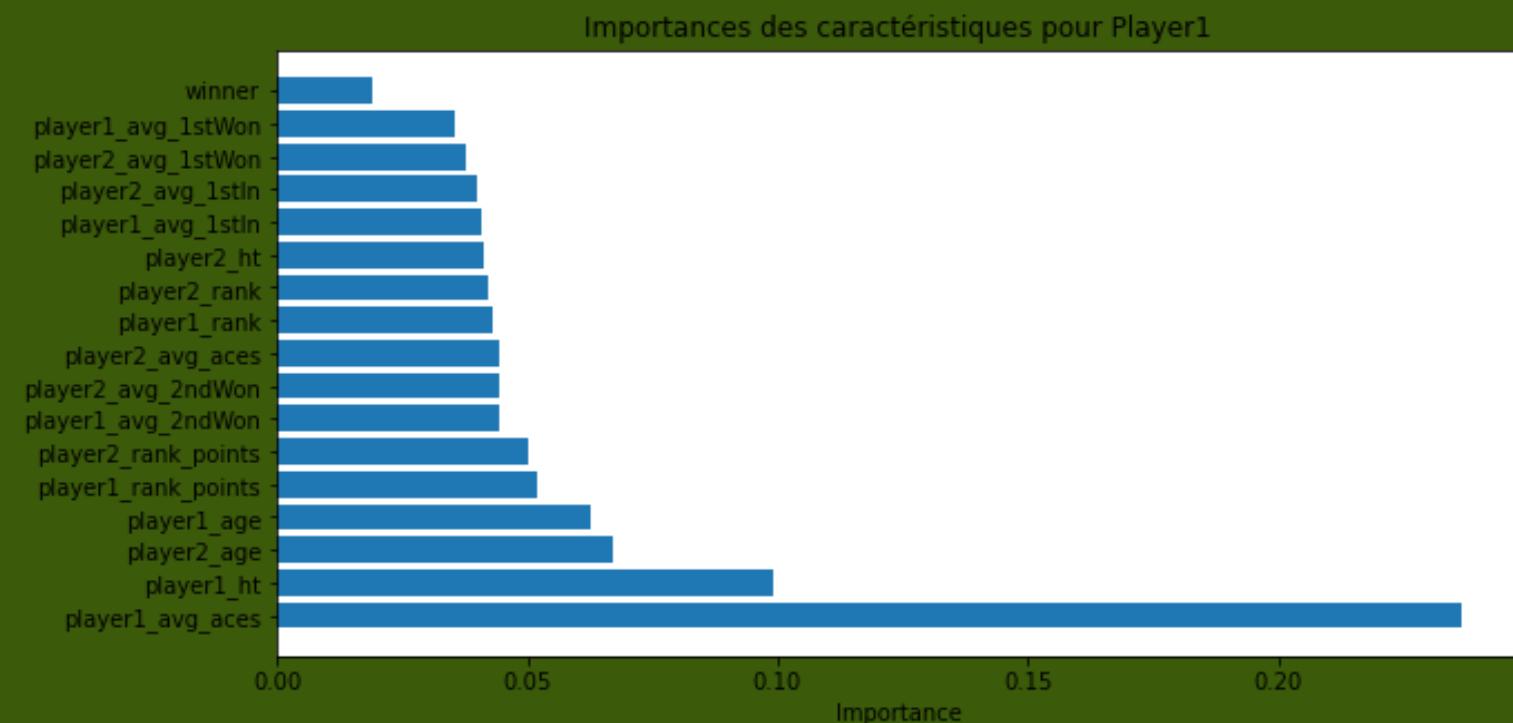
# Normalisation des nouvelles données avec le scaler déjà ajusté sur les données d'entraînement
# Use the scaler fitted on the selected features
new_match_scaled = scaler.transform(new_match_selected)

# Prédiction du nombre d'aces pour Player 1
predicted_aces = rf_model.predict(new_match_scaled)

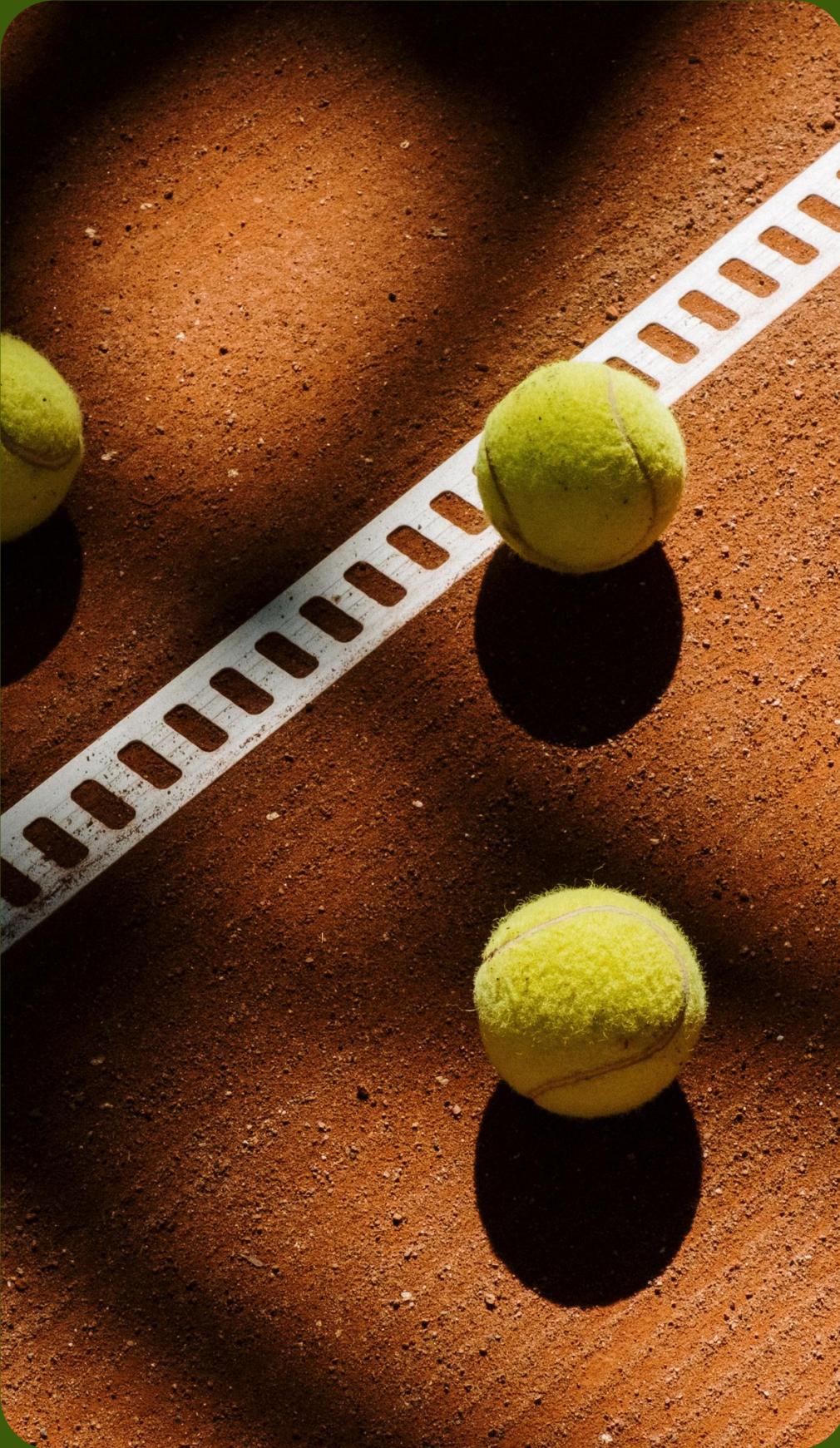
# Affichage de la prédiction
print(f"Nombre d'aces prédit pour Player 1 : {predicted_aces[0]:.2f}")

Nombre d'aces prédit pour Player 1 : 5.27
```

Comparaison avec d'autres modèles (SVM, XG boost,...) et d'autres variables prises en compte qui ne donnaient pas de meilleurs indicateurs métriques (**MAE ~ 2,3**)



L'exploitation des anciens matchs nous a permis de **créer une nouvelle base de données** sans avoir besoin de connaître le vainqueur en amont du match



PISTES D'AMÉLIORATION



Enrichissement des données : ajout des saisons précédentes, conditions météo, type de tournoi.



Optimisation des hyperparamètres pour des tests plus approfondis via **GridSearchCV**



Exploration **d'algorithmes plus avancés** comme **LSTM** pour capturer l'évolution temporelle des performances des joueurs

PROJET DATA

TENN'ACE



**MERCI POUR VOTRE
ATTENTION**