

Rapport de Projet Machine Learning

Premier League Players Statistics



Réalisé par:

Khalil Ayari

Siwar Saiidi

Rania Ben Hmida

Khalil Khedher

Nour Msaddak

Khalil Naddari

Amine Khalfaoui

Sommaire :

Chapitre 1 : Introduction	3
1.1 : Contexte général.....	3
1.2 : Problématique.....	3
1.3 : Objectifs du projet (DSO + BO).....	4
1.4 : Environnement de travail.....	5
Chapitre 2: Compréhension métier.....	6
2.1 : Domaine d'application.....	6
2.2 : Utilisateurs finaux et leur besoins.....	6
Chapitre 3: Compréhension des données.....	7
3.1 : Description de données.....	7
3.2 : Les variables et leur rôle.....	8
Chapitre 4 : Préparation des données.....	10
4.1 : Nettoyage des données.....	10
4.2 : Transformation des données.....	11
Chapitre 5 : Modélisation	13
5.1 : Modèles de machine learning utilisés.....	13
Chapitre 6 : Déploiement.....	17
Chapitre 7 : Conclusion et perspectives.....	21

Chapitre 1 : Introduction :

1.1 : Contexte général

Le football est l'un des sports les plus populaires au monde, attirant des millions de spectateurs et de passionnés chaque jour. Avec l'évolution des technologies et de l'analyse de données, de nombreuses équipes cherchent à améliorer leurs performances en exploitant les données des joueurs et des clubs. Dans ce contexte, notre projet s'inscrit dans une démarche innovante visant à utiliser l'apprentissage automatique (machine learning) pour analyser et prédire les résultats des matchs de football.

1.2 : Problématique

Le problème principal que nous cherchons à résoudre est de fournir des prédictions précises sur les résultats des matchs en se basant sur des statistiques complètes des joueurs et des clubs. En outre, ce projet vise à aider les équipes et les analystes à identifier les joueurs clés, à optimiser les compositions d'équipe, et à décider des transferts stratégiques.

1.3 : Objectifs du projet (DSO + BO)

Les objectifs de ce projet sont multiples :

- **Objectifs Stratégiques (BO) :**

Fournir des recommandations pour améliorer les performances des équipes et assister dans les décisions stratégiques comme les transferts ou les compositions d'équipe.

- **Objectifs Scientifiques et Opérationnels (DSO) :**

- Analyser les performances des joueurs et des clubs en fonction de divers indicateurs statistiques.
- Prédire les résultats des matchs de football avec précision.
- Développer une interface utilisateur intuitive pour la visualisation et l'utilisation des prédictions.

1.4 : Environnement de travail

Pour réaliser ce projet, nous avons utilisé les technologies suivantes :

- **Python** : pour le traitement des données et le développement des modèles de machine learning.
- **Flask** : pour la création de l'application web permettant le déploiement des modèles.
- **Pandas et NumPy** : pour la manipulation des données.
- **Scikit-learn** : pour la construction et l'évaluation des modèles prédictifs.

En exploitant un riche ensemble de données sur les clubs et les joueurs, ce projet contribue à l'intégration des technologies modernes dans le domaine du football et ouvre la voie à de nouvelles approches stratégiques dans ce sport.

Chapitre 2 : Compréhension métier

2.1 Domaine d'application:

L'application **Soccer Predictor** est une plateforme d'analyse avancée dédiée au monde du football. Son objectif principal est de fournir des insights basés sur les données pour prédire les positions des joueurs et identifier les talents clés. Cette application s'inscrit dans le contexte croissant de l'utilisation de la data science dans le sport, où les décisions basées sur les données deviennent essentielles pour améliorer les performances des équipes, maximiser le recrutement des talents et optimiser les stratégies de jeu.

2.2 Utilisateurs finaux et leurs besoins:

Utilisateur	Besoins spécifiques
Entraîneurs	<ul style="list-style-type: none">- Obtenir des prédictions fiables pour ajuster leurs formations.- Analyser les performances clés.
Recruteurs/Scouts	<ul style="list-style-type: none">- Accéder à un tableau comparatif des joueurs.- Identifier les meilleurs talents par position.
Analystes sportifs	<ul style="list-style-type: none">- Générer des rapports basés sur des indicateurs de performance précis.- Comprendre les tendances.
Fans	<ul style="list-style-type: none">- Explorer les statistiques des joueurs préférés.- Comparer les performances des stars du football.

Chapitre 3 : Compréhension des données

3.1:Description des données:

Les données utilisées dans ce projet contient des informations sur les performances individuelles de 481 joueurs de football. Ce fichier regroupe des statistiques clés permettant d'évaluer les contributions des joueurs à leurs équipes respectives.

Principales colonnes et leur description :

- **Name** : Nom complet du joueur.
- **Club** : Nom du club auquel appartient le joueur.
- **Position** : Poste occupé par le joueur sur le terrain (ex. : attaquant, milieu, défenseur, gardien).
- **Age** : Âge du joueur en années.
- **Goals per match** : Moyenne de buts marqués par match.
- **Performance Score** : Une évaluation numérique des performances globales du joueur basée sur divers indicateurs (ex. : buts, passes décisives, tacles réussis).
- **Nationality** : Nationalité du joueur.
- **Appearances** : Nombre total de matchs disputés par le joueur.

- **Minutes Played** : Durée totale de jeu en minutes.
- **Assists** : Nombre total de passes décisives réalisées.
- **Yellow Cards** : Nombre de cartons jaunes reçus.
- **Red Cards** : Nombre de cartons rouges reçus.

Ce fichier constitue une base riche et variée pour analyser les performances individuelles et leur impact sur les résultats des équipes.

3.2 : Les variables et leur rôle:

3.2.1. Variables descriptives :

Name, Club, Position, Age, Nationality : Ces variables fournissent des informations qualitatives sur les joueurs et permettent de filtrer ou de regrouper les données en fonction de différentes caractéristiques.

3.2.2. Variables de performance :

- **Goals per match, Performance Score, Appearances, Minutes Played, Assists** : Ces variables sont essentielles pour évaluer les

performances individuelles des joueurs. Elles serviront d'indicateurs clés dans l'analyse et les modèles prédictifs.

3.2.3.Variables explicatives :

- **Position, Age, Nationality, Club** : Ces variables influencent directement les performances des joueurs et permettent d'identifier des tendances ou des facteurs d'influence.

3.2.4.Variables cibles :

- Le résultat global des matchs ou les évaluations de performance des joueurs seront les variables à prédire dans le cadre de ce projet.
- Cette description claire et détaillée du fichier permet de mieux comprendre la structure et le rôle des données disponibles pour atteindre les objectifs du projet.

Chapitre 4 : Préparation des données

4.1 : Nettoyage des données:

Le nettoyage des données consiste à identifier et corriger les erreurs ou incohérences dans le jeu de données avant de l'utiliser dans le modèle de prédiction.

a. Traitement des valeurs manquantes :

	# Vérification des valeurs manquantes print(data.isnull().sum())																																																												
	<table><tr><td>Name</td><td>0</td></tr><tr><td>Jersey Number</td><td>8</td></tr><tr><td>Club</td><td>0</td></tr><tr><td>Position</td><td>0</td></tr><tr><td>Nationality</td><td>1</td></tr><tr><td>Age</td><td>1</td></tr><tr><td>Appearances</td><td>0</td></tr><tr><td>Wins</td><td>0</td></tr><tr><td>Losses</td><td>0</td></tr><tr><td>Goals</td><td>0</td></tr><tr><td>Goals per match</td><td>262</td></tr><tr><td>Headed goals</td><td>69</td></tr><tr><td>Goals with right foot</td><td>69</td></tr><tr><td>Goals with left foot</td><td>69</td></tr><tr><td>Penalties scored</td><td>262</td></tr><tr><td>Freekicks scored</td><td>262</td></tr><tr><td>Shots</td><td>262</td></tr><tr><td>Shots on target</td><td>262</td></tr><tr><td>Shooting accuracy %</td><td>262</td></tr><tr><td>Hit woodwork</td><td>69</td></tr><tr><td>Big chances missed</td><td>262</td></tr><tr><td>Clean sheets</td><td>309</td></tr><tr><td>Goals conceded</td><td>309</td></tr><tr><td>Tackles</td><td>69</td></tr><tr><td>Tackle success %</td><td>181</td></tr><tr><td>Last man tackles</td><td>378</td></tr><tr><td>Blocked shots</td><td>69</td></tr><tr><td>Interceptions</td><td>69</td></tr><tr><td>Clearances</td><td>69</td></tr><tr><td>Headed Clearance</td><td>69</td></tr></table>	Name	0	Jersey Number	8	Club	0	Position	0	Nationality	1	Age	1	Appearances	0	Wins	0	Losses	0	Goals	0	Goals per match	262	Headed goals	69	Goals with right foot	69	Goals with left foot	69	Penalties scored	262	Freekicks scored	262	Shots	262	Shots on target	262	Shooting accuracy %	262	Hit woodwork	69	Big chances missed	262	Clean sheets	309	Goals conceded	309	Tackles	69	Tackle success %	181	Last man tackles	378	Blocked shots	69	Interceptions	69	Clearances	69	Headed Clearance	69
Name	0																																																												
Jersey Number	8																																																												
Club	0																																																												
Position	0																																																												
Nationality	1																																																												
Age	1																																																												
Appearances	0																																																												
Wins	0																																																												
Losses	0																																																												
Goals	0																																																												
Goals per match	262																																																												
Headed goals	69																																																												
Goals with right foot	69																																																												
Goals with left foot	69																																																												
Penalties scored	262																																																												
Freekicks scored	262																																																												
Shots	262																																																												
Shots on target	262																																																												
Shooting accuracy %	262																																																												
Hit woodwork	69																																																												
Big chances missed	262																																																												
Clean sheets	309																																																												
Goals conceded	309																																																												
Tackles	69																																																												
Tackle success %	181																																																												
Last man tackles	378																																																												
Blocked shots	69																																																												
Interceptions	69																																																												
Clearances	69																																																												
Headed Clearance	69																																																												

```
# Step 1: Separate categorical and numeric columns
categorical_cols = data.select_dtypes(include=['object']).columns
numeric_cols = data.select_dtypes(include=['number']).columns

# Step 2: Impute categorical columns with the mode
for col in categorical_cols:
    mode_value = data[col].mode()[0] # Get the most frequent value
    data[col].fillna(mode_value, inplace=True) # Fill missing values with the mode

# Step 3: Impute numeric columns with the median (to handle outliers)
for col in numeric_cols:
    median_value = data[col].median() # Calculate the median
    data[col].fillna(median_value, inplace=True) # Fill missing values with the median

# Step 4: Verify Missing Data Handling
remaining_missing = data.isnull().sum() # Check for remaining missing values
if remaining_missing.sum() == 0: # Sum the series to check if all are handled
    print("\nAll missing values have been handled!")
else:
    print(f"\nRemaining missing values per column:\n{remaining_missing}")
```



All missing values have been handled!

b. Suppression des doublons:

```
[81] # Vérification des doublons
print(f"Duplicated rows: {data.duplicated().sum()}")
```



Duplicated rows: 0

```
[82] #supprimer les doublons
data.drop_duplicates(inplace=True)
```

4.2 : Transformation des données :

Une fois les données nettoyées, elles doivent être préparées pour être utilisées dans le modèle. Cela inclut la normalisation, l'encodage et d'autres transformations nécessaires.

a. Encodage des données catégoriques :

```
[104] def add_position_code_column(data_filtered, position):

    # Initialiser une liste pour stocker les codes
    position_codes = []

    # Parcourir chaque ligne pour déterminer le code de la position
    for pos in data_filtered[position]:
        if pos == "Midfielder":
            position_codes.append(0)
        elif pos == "Defender":
            position_codes.append(1)
        elif pos == "Forward":
            position_codes.append(2)
        elif pos == "Goalkeeper":
            position_codes.append(3)
        else:
            position_codes.append(-1) # Code par défaut pour les valeurs inconnues

    # Ajouter la nouvelle colonne au DataFrame
    data_filtered['PositionCode'] = position_codes

    return data_filtered

# Exemple d'utilisation
football = add_position_code_column(data_filtered, "Position")

# Affichage du résultat
print(football)
```

b. Normalisation - Standardisation :

```
# Normalisation des colonnes numériques
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
features = ['Goals', 'Assists', 'Shots on target', 'Big chances created', 'Passes', 'Tackles']
data[features] = scaler.fit_transform(data[features])
```

c. Séparation des données de train et de test :

```
from sklearn.model_selection import train_test_split
df = data_filtered[features + [target]].dropna()
X = df[features]
y = df[target]
```

```
# Split the data into train and test set
X_scaled = scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, random_state=42, test_size=0.2)
```

Chapitre 5 : Modélisation

5.1 : Modèles de machine learning utilisés

1. KNN (K-Nearest Neighbors):

Le modèle KNN est un algorithme supervisé utilisé pour la classification. Il fonctionne en calculant la distance entre les points de données.

```
[43] print( "The best accuracy was with", mean_acc.max(), "with k=", mean_acc.argmax()+1)
```

➡ The best accuracy was with 0.7010309278350515 with k= 5

2. Random Forest:

Random Forest est un ensemble de plusieurs arbres de décision qui vote pour prédire les classes.

```
✓ [53] # Evaluate  
0s print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
```

➡ Random Forest Accuracy: 0.865979381443299

3. Support Vector Machines (SVM)

SVM est un algorithme puissant qui cherche un hyperplan optimal pour séparer les classes.

```

➡ Train set Accuracy with linear kernel: 0.5416666666666666
Test set Accuracy with linear kernel: 0.4742268041237113
Train set Accuracy with poly kernel: 0.4921875
Test set Accuracy with poly kernel: 0.38144329896907214
Train set Accuracy with rbf kernel: 0.6119791666666666
Test set Accuracy with rbf kernel: 0.4639175257731959
Train set Accuracy with sigmoid kernel: 0.4453125
Test set Accuracy with sigmoid kernel: 0.3711340206185567
The best accuracy was with 0.47 using the 'linear' kernel.

```

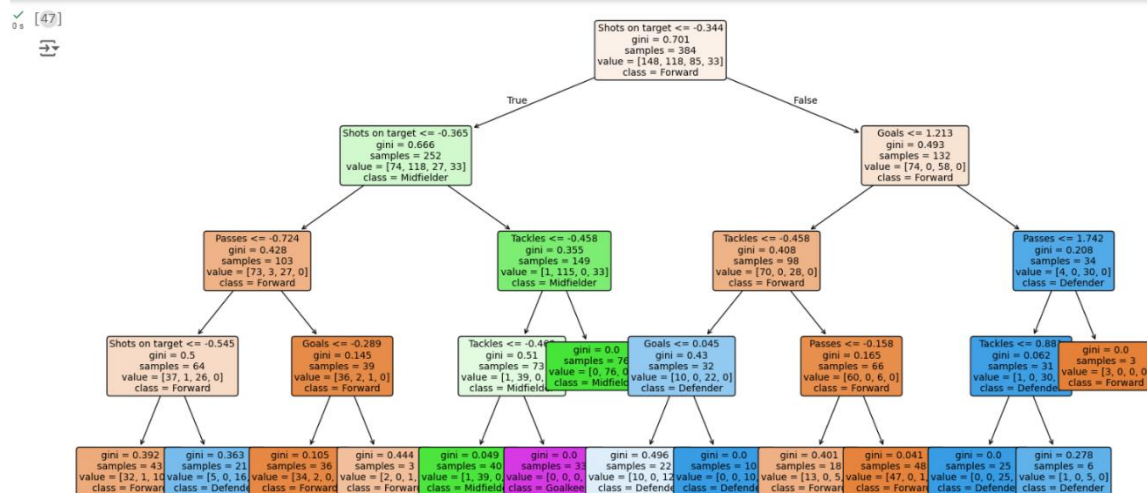
4. Arbre de Décision (ADD):

Un arbre de décision utilise une structure hiérarchique pour séparer les données en fonction de critères successifs.

```

➡ L'accuracy de l'arbre de décision est : 0.88 with depth = 4

```



5. XGBoost :

XGBoost est un algorithme de gradient boosting rapide et performant.

```
✓ [50] # Evaluate  
js print("XGBoost Accuracy:", accuracy_score(y_test, y_pred))
```

⇒ XGBoost Accuracy: 0.845360824742268

6. K-means (Clustering) :

K-means est un algorithme non supervisé utilisé pour regrouper les données.

```
✓ [72] labels = k_means.labels_  
0 s print(labels)
```

```
⇒ [3 2 3 1 1 2 2 2 1 3 0 1 1 1 3 1 3 2 3 3 1 1 3 3 3 3 2 1 3 3 3 2 3 1 3 3 2  
3 1 3 3 3 3 1 3 3 3 3 3 3 3 3 1 3 3 2 1 3 3 1 3 3 1 2 1 1 3 3 3 3 3 3 3  
1 3 2 1 1 2 2 2 3 3 0 3 3 3 1 2 3 1 1 3 1 3 2 2 1 2 3 0 3 3 1 2 2 2 3 1 3  
3 3 3 3 3 1 3 1 3 2 2 2 2 2 0 1 3 2 3 2 1 1 2 0 3 3 1 1 2 1 3 3 1 1 1 2 1  
1 2 3 3 2 2 1 3 1 3 3 3 2 3 3 1 1 3 2 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 2 3 3 3 1 1 0 3 1 1 3 3  
1 1 1 2 3 3 3 3 2 3 3 3 1 3 3 1 3 3 1 1 2 1 2 2 2 3 1 2 0 3 0 1 3 3 3 3 1  
3 2 2 1 3 3 3 3 1 1 2 0 2 1 1 1 3 2 3 3 2 1 2 1 0 3 3 3 2 3 2 2 2 3 3 3 1  
3 2 2 2 1 3 2 3 1 3 3 2 0 1 3 1 3 3 0 1 3 1 1 3 3 1 3 2 1 3 1 3 2 1 1 3 3  
2 1 1 3 3 3 3 1 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 0 3 3 3 1 3 3 1 3 3 3 3 3 1  
3 1 3 1 2 3 1 3 3 3 2 2 3 3 2 3 3 1 1 3 3 2 3 2 2 1 3 1 2 3 1 1 1 2 2 3 2  
3 3 2 3 3 1 1 3 1 3 3 3 2 3 3 3 3 3 2 2 3 3 3 3 3 3 3 3 3 3 3 2 3 3 2 3  
1 1 3 1 3 1 1 1 0 3 2 1 3 1 3 3 3 3 3 1 1 1 1 3 3 3 3 1 3 1 1 1 3 3 1 3 3]
```

```
✓ [73] # Get the cluster centers  
0 s centers = k_means.cluster_centers_  
print("Cluster Centers:\n", centers)
```

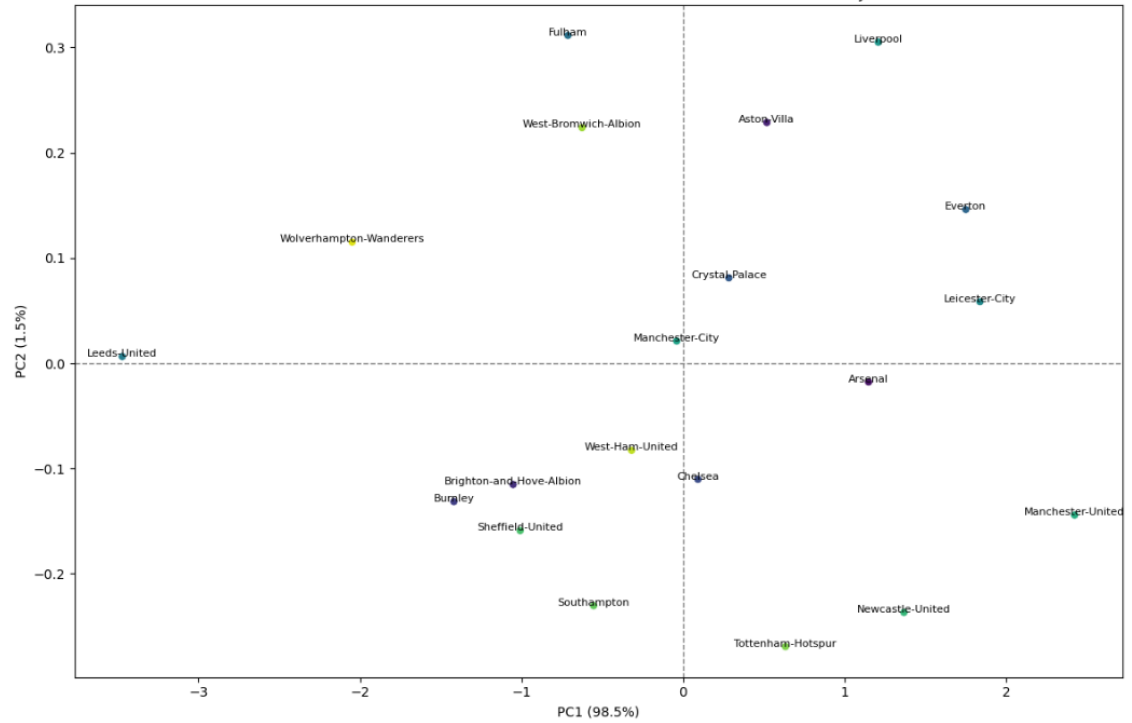
```
⇒ Cluster Centers:  
[[2.36428571e+01 2.96428571e+01 5.97142857e+01 2.82142857e+01  
1.39837143e+04 5.59000000e+02]  
[1.31774194e+01 7.83064516e+00 4.21774194e+01 1.16854839e+01  
3.11137903e+03 1.22895161e+02]  
[1.83703704e+01 1.45185185e+01 5.35432099e+01 2.11111111e+01  
7.15965432e+03 2.71111111e+02]  
[2.79007634e+00 1.74045802e+00 1.30381679e+01 2.61832061e+00  
6.00480916e+02 2.97175573e+01]]
```

7. PCA (Analyse en Composantes Principales) :

PCA est une méthode de réduction de dimensionnalité utilisée pour simplifier les données.

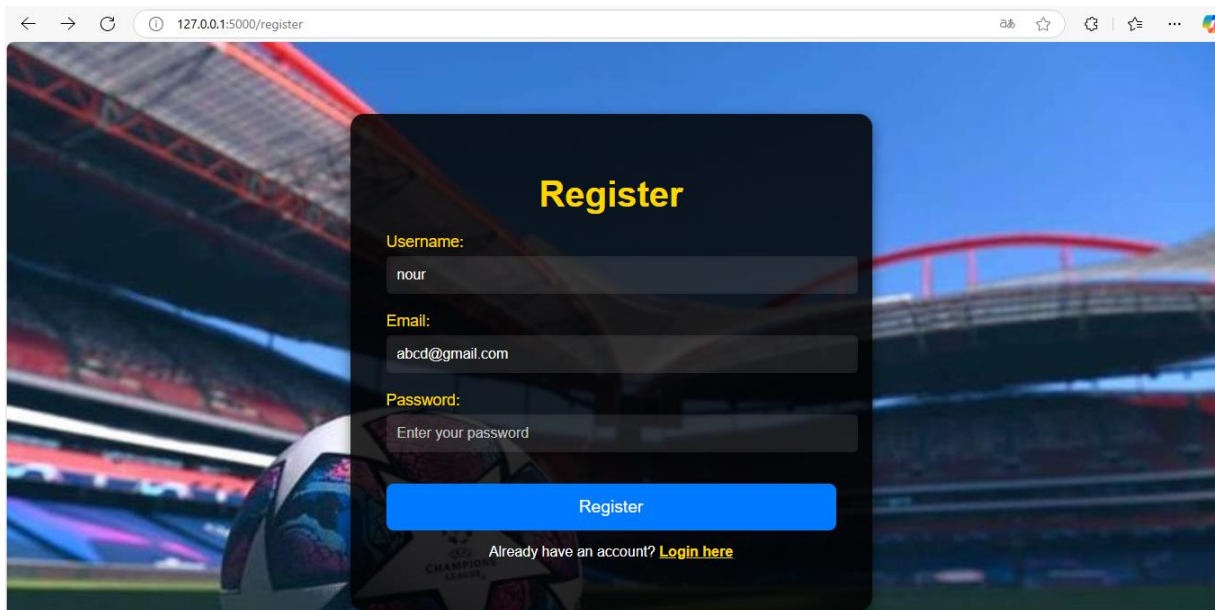


Carte des Individus des Clubs Basée sur la Performance et le Nombre de Joueurs



Chapitre 6 : Déploiement

Le déploiement est une étape clé dans le cycle de vie d'un projet de machine learning. Cela implique de rendre votre modèle et votre application accessibles aux utilisateurs finaux via une interface fonctionnelle. Voici comment appliquer cette étape à votre projet de **Soccer Predictor**.



127.0.0.1:5000/register

Register

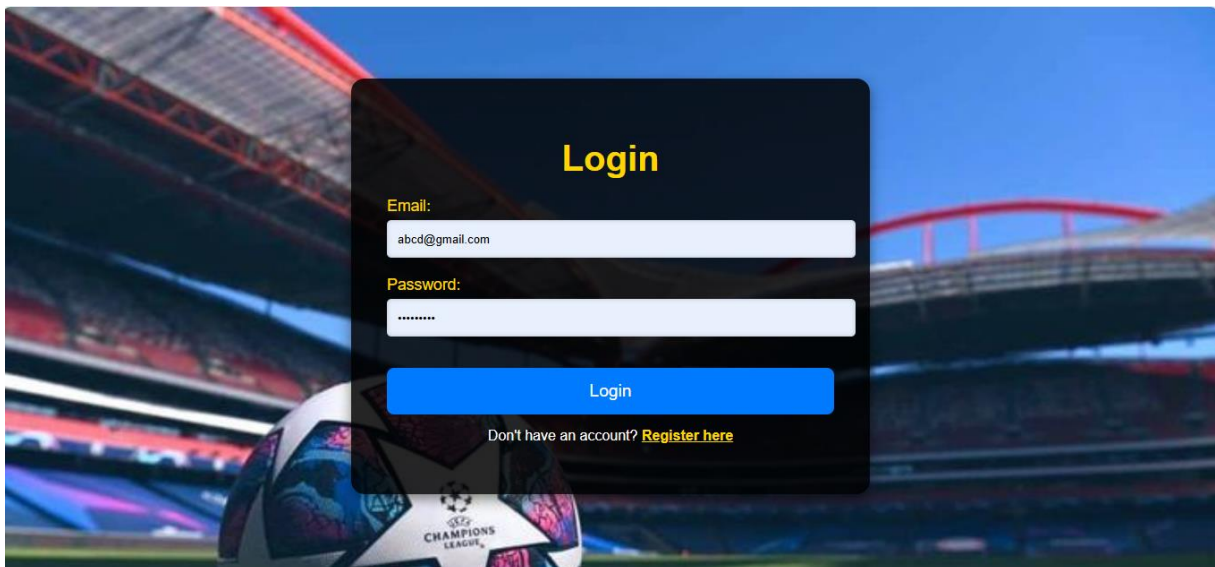
Username:
nour

Email:
abcd@gmail.com

Password:
Enter your password

Register

Already have an account? [Login here](#)



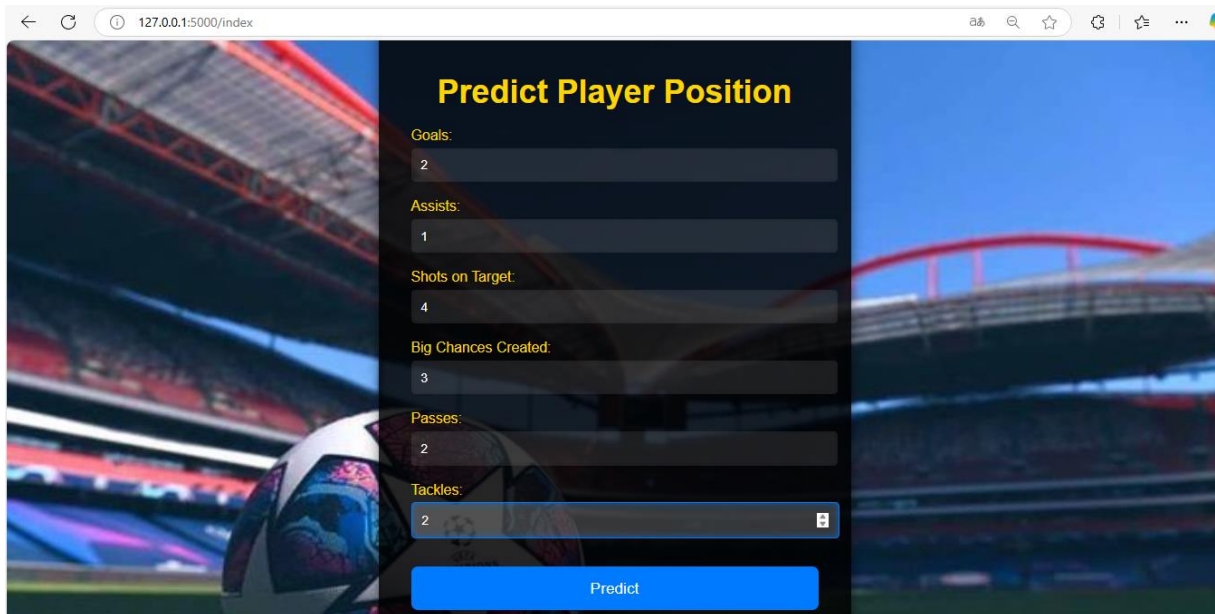
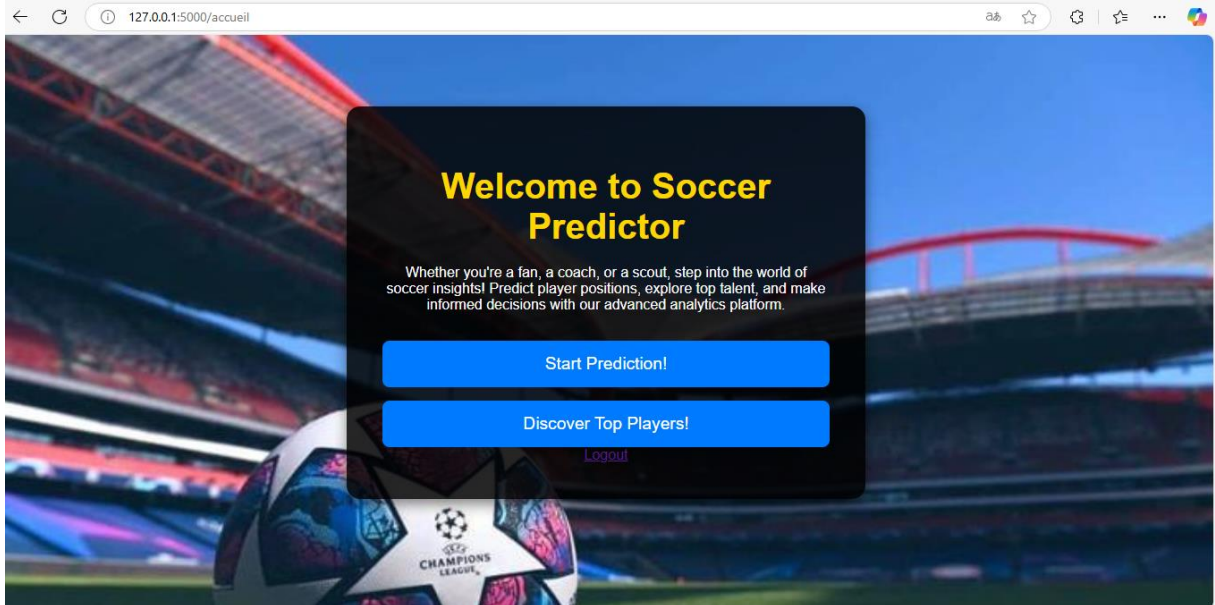
Login

Email:
abcd@gmail.com

Password:

Login

Don't have an account? [Register here](#)



127.0.0.1:5000/predict

Prediction Result

Predicted Position: **Forward**

Similar Players

Name	Club	Position	Goals	Assists
Callum Robinson	West-Bromwich-Albion	Forward	1	1
James Rodríguez	Everton	Midfielder	1	1
Ivan Cavaleiro	Fulham	Forward	3	1
Daniel Podence	Wolverhampton-Wanderers	Midfielder	1	2
Bertrand Traoré	Aston-Villa	Forward	2	1

[Back to Home](#)

127.0.0.1:5000/recommend

Top 5 Players Recommendation

Select Position:
Midfielder

[Get Top Players](#)

Name	Club	Position	Performance Score
Mesut Özil	Arsenal	Midfielder	4.0
Lucas Torreira	Arsenal	Midfielder	4.0
Ainsley Maitland-Niles	Arsenal	Midfielder	4.0
Mohamed Elneny	Arsenal	Midfielder	4.0
Joseph Willock	Arsenal	Midfielder	4.0

[Back to Home](#)

Chapitre 7 : Conclusion et Perspectives:

Ce projet de machine learning appliqué au domaine du football a permis d'atteindre des résultats significatifs. Les modèles développés ont montré une capacité à analyser les performances des équipes et à prédire les résultats des matchs en s'appuyant sur des données pertinentes. Cette solution offre une aide précieuse dans la prise de décision, notamment pour les entraîneurs et analystes sportifs.

Cependant, certaines limites subsistent. La qualité et la diversité des données disponibles peuvent impacter la fiabilité des prédictions, tandis que l'utilisation d'un unique modèle limite les comparaisons. De plus, l'absence de variables contextuelles telles que les conditions météo ou l'état psychologique des joueurs réduit la profondeur des analyses.

Pour aller plus loin, plusieurs améliorations sont envisageables. Il serait pertinent d'intégrer des données plus complètes et actualisées, d'explorer d'autres algorithmes de machine learning pour optimiser les performances, et de développer une application interactive permettant des analyses en temps réel. Ces évolutions renforceraient l'efficacité et l'applicabilité de la solution dans le domaine sportif.