Documentation du projet axé sur les performances des équipes et des joueurs durant la saison 25/26 au sein du Big 5

Introduction au projet

L'objectif de cette application web sera de visualiser les performances des équipes ainsi que des joueurs sur la saison 25/26. Les données des joueurs proviennent de Opta The Analyst, Fbref et Transfermarkt. Cette analyse l'analyse portera ainsi sur la saison 25/26 pour les compétitions suivantes : Ligue 1, Bundesliga, Premier League, La Liga, Serie A**.** Concrètement, cela vous permettra notamment d'évaluer le style de jeu de chaque équipe et de repérer le joueur de votre choix en fonction de ses performances statistiques au cours de la saison en cours.

Extraction des données / Pré-traitement des données

- Équipe

De par les besoins en données sur les performances des équipes durant la saison en cours, nous avons collecté des informations provenant de divers fournisseurs de données.

Concernant **Opta**, nous avons scrappé les données des cinq grands championnats en récupérant les tables suivantes : ***Attaque, Défense, Pressing, Séquence et Autre***. Toutes les statistiques de ces tables sont récupérées (sauf celles qui sont redondantes), et renommées avec comme préfixe le nom de leur table respective. Par ailleurs, ces données sont **actualisées hebdomadairement** via GitHub Actions, et ceux pour chaque championnat.

Ensuite, nous avons structuré notre base de données à partir des statistiques disponibles sur **FBref**, qui propose un large éventail d'indicateurs avancés couvrant aussi bien les phases offensives que défensives, la possession, le pressing ou encore les actions du gardien de but. FBref organise ses données en différentes catégories : ***Standard Stats, Shooting, Passing, Pass Types, Goal and Shot Creation, Defensive Actions, Possession, Playing Time, Miscellaneous Stats*****,** entre autres.

Afin d'intégrer ces informations dans notre modèle, nous avons d'abord récupéré les données des équipes grâce au scraping en Python. Historiquement, l'ensemble de cette collecte était automatisée, mais nous utilisons désormais une extraction manuelle assistée par un script Python pour garantir une meilleure maîtrise du nettoyage et de la structuration des données. Ce processus inclut notamment une phase de nettoyage, au cours de laquelle nous avons corrigé les noms des ligues, harmonisé différentes variables textuelles et supprimé des colonnes redondantes telles que *Rk*, *Squad*, *Comp*, etc.

À cela s'ajoutent également les données salariales des équipes, intégrées afin d'apporter une dimension économique à l'analyse, ainsi que la position de chaque équipe dans le classement de son championnat.

Nous avons ensuite joint ces variables par équipe grâce à un **mapping basé sur les noms d'équipes**, afin d'obtenir une base de données complète. Nous avons également créé plusieurs variables issues des informations obtenues précédemment, telles que le classement des équipes

dans leur championnat, le pourcentage d'attaques selon le style de jeu pour chaque équipe, ou encore un ajustement des métriques défensives en fonction de la possession de balle de l'équipe donnée.

À partir de ce fichier, nous avons établi un système de notation par catégorie en fonction des statistiques disponibles. Nous pouvons les séparer en quatre groupes distincts :

- **Statistiques avec le ballon** : Occasions générées, Finition, Coups de pied arrêtés (offensifs), Construction d'occasion, Projection, Centres, Dribbles.
- **Statistiques sans le ballon** : Occasions concédées, Actions défensives, Coups de pied arrêtés (défensifs), Efficacité du gardien.
- **Statistiques liées au style de jeu** : Jeu direct, Contre-attaque, Possession, Pressing.
- **Autres** : Classement au championnat, Duels au sol, Duels aériens, Fautes provoquées, Fautes commises, Pertes de balle, Remplacements.

Chacune de ces **sous-catégories** contient plusieurs statistiques associées, accompagnées d'un **coefficient** en fonction de leur importance pour catégoriser un groupe. À noter que le choix de ces métriques et coefficients a été réalisé en concertation entre les deux membres du projet et reste donc entièrement subjectif.

Dans la même logique, chaque catégorie de statistiques possède un coefficient permettant d'obtenir une évaluation globale de chaque équipe, comprise entre 0 et 100. Cette évaluation détermine le classement pour le power ranking au sein des 96 équipes du Big 5. Enfin, nous avons tenu à prendre en compte les forces de chaque championnat en appliquant une **légère pénalité fondée sur l'indice de puissance des championnats** calculé par Opta en mai 2025.

- Joueur

Pour l'analyse individuelle des joueurs, nous avons construit notre base de données à partir du jeu de données disponible sur **Kaggle** :
https://www.kaggle.com/datasets/hubertsidorowicz/football-players-stats-2025-2026.

Auparavant, ces informations étaient également obtenues par scraping, mais ce processus n'est plus utilisé. Nous nous appuyons désormais directement sur ce fichier, ce qui permet d'assurer une meilleure cohérence et une réduction du bruit dans les données initiales.

Une phase de nettoyage approfondi a ensuite été réalisée afin d'harmoniser et fiabiliser les variables. Parmi les principales opérations effectuées :

- **Standardisation des positions** : les abréviations ont été remplacées par leur forme complète, de façon à obtenir des catégories plus explicites, par exemple :
    - *DF → Defender*
    - *MF → Midfielder*
    - *FW → Forward*
    - *GK → Goalkeeper*
- **Correction et uniformisation des noms de pays**, remplacés par leur version complète plutôt que des formes abrégées.

- **Élimination des doublons**, permettant de conserver un seul enregistrement par joueur.

En complément, nous avons intégré des **données salariales** provenant de **Capology**, enrichissant ainsi la base de données avec une dimension économique essentielle pour l'analyse de performance et de valeur des joueurs

Pour compléter notre base d'informations, nous avons également intégré des données issues de **Transfermarkt**, en utilisant l'API non officielle disponible à l'adresse suivante : **https://github.com/felipeall/transfermarkt-api**.

Conformément aux instructions fournies, nous avons installé **Docker** puis configuré l'environnement nécessaire afin d'effectuer des requêtes vers l'API. Ce processus nous permet de récupérer, dans un premier temps, la liste des clubs, puis les joueurs appartenant à chacun d'eux. À partir de ces informations, nous effectuons ensuite des requêtes supplémentaires pour obtenir des données détaillées sur chaque joueur.

Les informations fournies par Transfermarkt sont particulièrement riches et apportent une précision que d'autres sources ne proposent pas. Parmi les variables les plus importantes que nous intégrons à notre modèle, on retrouve notamment :

- La **position naturelle** du joueur, décrite de manière granulaire (par exemple *Left Winger*, *Right Back*, *Attacking Midfielder*, etc.), plutôt qu'une simple catégorie générale.
- Les **données de représentation** (agent ou structure qui gère le joueur).
- L'historique et la situation actuelle concernant les **blessures**.
- Un ensemble d'informations contextuelles comprenant :
    - le nom du joueur *(player_name)*
    - le poste du joueur *(position)*
    - sa date de naissance *(dateOfBirth)*
    - son âge *(age)*
    - sa nationalité *(nationality)*
    - son club *(currentClub)*
    - *sa taille (height)*
    - son pied fort *(foot)*
    - son année d'arrivée dans son club actuel *(joinedOn / joined)*
    - son dernier club (*signedFrom)*
    - la date de fin de contrat *(contract)*
    - sa valeur sur le marché *(marketValue)*
    - *status* (indiquant si le joueur est blessé, suspendu ou capitaine)

Ensuite, à partir de ces données, nous avons **agrégé certaines statistiques par 90 minutes et ajusté d'autres en fonction de la possession de leur équipe.** Le but derrière l'ajustement à la possession est d'analyser les joueurs de façon objective concernant les actions défensives comme si l'équipe du joueur en question avait le ballon 50% du temps. Nous avons ensuite effectué une **jointure entre les deux fournisseurs de données**. Pour cela, nous avons utilisé plusieurs stratégies afin d'englober un maximum de joueurs sans commettre d'erreurs.

La première étape a consisté à associer les joueurs ayant le même nom, le même championnat, la même année de naissance, ainsi que le même poste lorsque le joueur était un gardien (le seul poste où les fournisseurs ne présentent aucune différence). Cette phase a permis d'associer la majorité des joueurs, mais certaines variations dans les noms entre les deux fournisseurs nous ont amenés à aller plus loin. Nous avons donc utilisé une fonction de similarité textuelle (fuzzy matching) selon la logique suivante : si le pourcentage de similarité est suffisant, combiné au même championnat, à la même année de naissance et au même poste (pour les gardiens), alors nous considérons qu'il s'agit bien du même joueur.

Enfin, nous avons procédé à une association manuelle pour les joueurs dont les noms étaient trop éloignés entre les deux fournisseurs, mais dont les autres informations correspondaient parfaitement. À noter que plus le pourcentage de correspondance sur le nom est faible, plus le risque d'erreur est élevé. Nous avons donc décidé de ne pas descendre en dessous d'un seuil de 65 %, afin de garantir une association fiable.

Nous avons ensuite établi un **système de notation des joueurs**, compris entre 0 et 100 (par catégorie et au global), en fonction des performances des autres joueurs évoluant au même poste (et dans la même catégorie de poste). Une note élevée dans une catégorie, ou de manière générale, reflète un haut niveau de performance sur la saison 2025/2026. Par souci de cohérence, certaines variables ont été inversées (erreurs, pertes de balle, etc.). De plus, chaque poste possède des coefficients adaptés aux exigences spécifiques du rôle. Par exemple, un attaquant performant doit présenter de bonnes statistiques en finition et en création d'occasions pour obtenir une bonne note. De la même façon, chaque catégorie de statistiques dispose d'un poids destiné à valoriser les joueurs obtenant de bons résultats dans les dimensions clés associées. À noter que les gardiens disposent de leur propre système de notation, en raison de la particularité de leur poste par rapport aux joueurs de champ. Enfin, des **pénalités peuvent être appliquées à la note globale** en fonction de la **puissance du championnat** dans lequel évolue le joueur et du **pourcentage de minutes disputées** sur l'ensemble de la saison.

Liste des pages

L'application sera composée en 2 parties : l'analyse des équipes et l'analyse des joueurs.

Pages dédiées aux équipes (6 pages)

- **Accueil** : Présentation du projet et des sources de données.
- **Analyse d'une équipe** : Exploration détaillée de l'équipe choisie à travers différentes statistiques.
- **Comparaison entre équipes** : Mise en parallèle de deux équipes pour une analyse comparative.
- **Classement – Statistiques de base** : Classement des équipes selon une statistique simple sélectionnée.
- **Classement – Statistiques avancées** : Classement des équipes selon une statistique avancée choisie.
- **Power Ranking** : Classement global des 96 équipes du Big 5, ainsi qu'un classement par championnat.

Pages dédiées aux joueurs (6 pages)

- **Accueil** : Présentation du projet et des sources de données.
- **Analyse d'un joueur** : Analyse détaillée du joueur sélectionné à travers différentes statistiques.
- **Comparaison entre joueurs** : Comparaison de deux joueurs évoluant au même poste.
- **Classement – Statistiques de base** : Classement des joueurs selon une statistique simple choisie.
- **Classement – Statistiques avancées** : Classement des joueurs selon une statistique avancée.
- **Scouting** : Recherche de joueurs correspondant aux critères définis par l'utilisateur (informations générales, statistiques de base, statistiques avancées).

Arborescence du projet pour la mise en place de l'application

Application
├── scripts (team, player, big_5_performance.py) : Toutes les scripts du projet par type
├── data (team, player) : Toutes les données du projet regroupées par type
├── image : Images utilisées pour le projet
├── README.md
├── .github : Tous les workflows utilisés pour l'automatisation des données
└── requirements.txt : Fichier de dépendances de librairies

Mise en place de l'application

Il est important de préciser que l'application contiendra 3 versions : Français, Anglais et Espagnol.

Accueil

Comme expliqué précédemment, la partie de l'accueil présentera brièvement les composantes du projet, et donne accès à diverses ressources (Documentation, Code source du projet).

Section analyse

En tête du projet

Nous créerons au préalable plusieurs fonctions (construction des tableaux d'analyse, glossaire, traductions, etc.) ainsi que diverses variables (glossaire, listes de données par catégorie) afin d'éviter de surcharger la longueur du projet, compte tenu des trois langues disponibles et des types d'analyses présentant de nombreuses similarités.

Affichage de l'application

- **Équipe**

Pour l'**analyse des équipes**, l'utilisateur devra d'abord sélectionner le championnat de son choix afin de choisir l'équipe qu'il souhaite étudier. L'analyse sera structurée en plusieurs

parties : présentation des informations générales sur l'équipe (nom, championnat, classement, power ranking, etc.), affichage du top 5 de ses joueurs sur la saison selon nos critères de performance, liste des cinq joueurs les mieux valorisés sur Transfermarkt, ainsi que plusieurs graphiques comparatifs (offensifs, défensifs, style de jeu, autres) permettant de situer l'équipe par rapport au Big 5 ou à son championnat. Une liste de cinq équipes similaires sera également proposée à partir des statistiques disponibles.

La section **Duel** offrira une comparaison entre deux équipes choisies par l'utilisateur, en confrontant leurs statistiques pour faciliter l'analyse comparative.

La section **Stats +** permettra d'obtenir un classement fondé sur une statistique agrégée par catégorie au choix (actions créées, finition, dribbles, etc.), avec la possibilité de filtrer par championnat.

La section **Stat** suivra la même logique, mais appliquée à l'ensemble des statistiques brutes disponibles (xG, PPDA, nombre de contre-attaques par 90 minutes, etc.). L'utilisateur devra au préalable sélectionner la catégorie de la statistique souhaitée afin de faciliter la navigation. Un glossaire sera également accessible pour expliciter chaque statistique.

Enfin, la section **Power Ranking** permettra d'afficher facilement le classement des 96 équipes, avec un filtrage possible par championnat, et de comparer rapidement leurs statistiques respectives.

- **Joueur**

Ensuite, pour l'**analyse du joueur**, premièrement, on demandera à l'utilisateur de choisir le joueur de son choix. Cette page contiendra le profil du joueur avec ses informations de base (Nom, Photo, Poste, Club ect...), son radar statistique, et une table des 5 joueurs le ressemblant le plus statistiquement. À noter que le radar peut être adaptable selon le pays, championnat, tranche d'âge. Les statistiques affichées sur le radar sont celles semblant d'intérêt par rapport au poste du joueur (en bleu), et sera comparé avec la médiane du groupe choisie (en rouge). Par ailleurs, un glossaire des statistiques sera disponible, et dépliable si besoin.

La page de **comparaison de joueur** suivra la même logique avec la possibilité de choisir deux joueurs (du même poste) que l'utilisateur souhaite analyser. Leur profil sera affiché côte à côte, suivi de leur radar respectif (en bleu et rouge).

Concernant la page de **classement des statistiques de base**, l'utilisateur pourra obtenir un classement selon la catégorie de son choix (Finition, Dribble, ect..) . Il pourra par ailleurs le filtrer selon le poste, tranche d'âge, pays ect.. Le podium des meilleurs joueurs selon ces critères sera affiché avec les informations générales sur l'ensemble des joueurs répondant aux requêtes de l'utilisateur.

Dans la même idée, à propos de la page de **classement des statistiques avancées**, il sera demandé à l'utilisateur la statistique qu'il souhaite analyser. À partir de ce choix, un classement des meilleurs joueurs selon cette métrique sera affiché avec un podium pour les 3 premiers, suivie par la suite du classement sous forme de tableau avec les informations de base sur ces derniers. La sidebar contiendra notamment des filtres facultatifs (Poste, Club, Championnat, Tranche d'Age, Valeur sur le Marché), ainsi qu'un glossaire des statistiques. Par ailleurs, une

image sera affichée avant le choix de l'utilisateur, afin de ne pas laisser vide la page (comme cela est le cas pour chaque page).

Enfin la page de **Scout** permettra d'obtenir une liste de joueurs selon les informations souhaitées par l'utilisateur (Informations de base, Statistiques de Base et Avancée). Un récapitulatif des choix effectués sera disponible en sidebar, ainsi qu'un glossaire des statistiques.