Documentation du projet Application web

Introduction au projet

L'objectif de ce projet est de poursuivre le travail effectué lors de mon mémoire de M1 : Analyse comparative de 3 facteurs de performance dans le football : l'impact du 1er but, la distribution temporelle des buts et l'influence de l'avantage du terrain sur le match (domicile/extérieur) entre les équipes de jeunes (U17N et U19N) et le monde professionnel (Ligue 1).

Afin d'étendre cette analyse, on comparera cette fois ci les compétition suivantes, et ceux des saisons 2021/2022 à 2024/2025 (lorsque cela est possible) :

- **Ligue 1**
- **Ligue 2**
- **National 1**
- **National 2**
- **Championnat U19N**
- **D1 Féminine**
- **D2 Féminine**

Pour rappel, on s'occupera des facteurs suivants :

- **l'influence du 1er but sur le match**
- **la distribution temporelles des buts**
- **l'influence du paramètre domicile/exterieur**
- **l'avantage du terrain x 1er but**
- **le nombre de buts par match**

Ce travail s'articulera est plusieurs étapes avec :

- **la récupèration des données via l'utilisation du web scrapping auprès du site Sofa Score**
- **le stockage de ces données dans des tables via l'utilisation de Supabase**
- **l'analyse des données collectées**
- **la mise en page de l'application web**
- **le déploiement de cette application**.

Web scraping / Stockage des données sur Supabase

Import des librairies

On va ensuite importer des librairies qui seront utile pour plus tard tels que pandas, numpy ou encore BeautifulSoup. À noter qu'il faudra appliquer la ligne suivante au sein de votre terminal afin d'installer toutes les dépendances nécessaires au bon fonctionnement du projet : pip install -r requirements.txt

Liaison avec la base de données Supabase

On se servira de variables d'environnement afin de stocker nos données personnels (url du projet et la clé api associé) pour accéder au projet sur Supabase.

Création des requêtes permettant la création des tables sur Supabase (à faire sur Supabase)

Veuillez créer également ces tables au sein de votre projet directement sur Supabase. On ajoutera les données plus tard dans le projet.

- **Création de la table Compétition**

CREATE TABLE Competition (id_competition SERIAL PRIMARY KEY,competition_name VARCHAR(255) NOT NULL,country_name VARCHAR(255) NOT NULL,link_url TEXT NOT NULL);

- **Création de la table Saison**

CREATE TABLE Season (id_season SERIAL PRIMARY KEY,season_name VARCHAR(255) NOT NULL,id_competition INT NOT NULL,link_url TEXT NOT NULL,CONSTRAINT fk_competition FOREIGN KEY (id_competition) REFERENCES Competition(id_competition) ON DELETE CASCADE);

- **Création de la table Équipe**

CREATE TABLE team (id_team INT PRIMARY KEY,team_name TEXT NOT NULL);

- **Création de la table des informations des matchs**

CREATE TABLE info_match (id_match INT PRIMARY KEY,id_season INT NOT NULL,id_home_team INT NOT NULL,id_away_team INT NOT NULL,match_date DATE NOT NULL,link_url TEXT,FOREIGN KEY (id_home_team) REFERENCES team(id_team),FOREIGN KEY (id_away_team) REFERENCES team(id_team));

- **Création de la table des informations des buts sur les matchs**

CREATE TABLE Info_goal (id_match INT NOT NULL,score_home INT,score_away INT,result INT,home_0_15 INT,away_0_15 INT,home_16_30 INT,away_16_30 INT,home_31_45 INT,away_31_45 INT,home_46_60 INT,away_46_60 INT,home_61_75 INT,away_61_75 INT,home_76_90 INT,away_76_90 INT,PRIMARY KEY (id_match),FOREIGN KEY (id_match) REFERENCES info_match(id_match));

Stocker les informations des compétitions française

Ensuite, on va créer la classe Compétition, et sa fonction permettant l'insertion de ces données dans la table sur Supabase associé (et créé précedemment)

Pour récupérer les données de compétitions, on va effectuer du web scrapping auprès du site Sofa Score et le lien ci dessous :

- https://www.sofascore.com/fr/football/france.

On va s'infllltrer dans la page source afin de récupérer les informations de chaque compétition via la librairie BeautifulSoup. Après avoir ciblé la classe contenant ces informations, on stockera les données suivantes :

- l'identifiant de la compétition
- le nom de la compétition
- le nom du pays de la compétition
- le lien url de la compétition.

Par la suite, ces dernières seront stocker dans un dataframe, puis inserer dans la table associé, via la fonction insert_competition créé précedemment.

Stocker les informations des saisons française

Par la suite, avec l'aide de la librairie psycopg2, on va accéder à la table précedemment créé, dans le but de stocker les informations des saisons disponible via la table des compétitions. On utilisera la librairie conn pour effectuer la requête suivante : SELECT id_competition, link_url FROM competition. Cette requête nous ressort ainsi les identifiants et les liens urls de chacune des compétitions françaises disponibles sur le site Sofascore.

Dans la même logique que la section précedente, on va créer la classe Season et sa fonction associé. Cette classe contiendra les informations suivantes :

- **l'identifiant de la saison**
- **le nom de la saison**
- **l'identifiant de la compétition**
- **le lien url de la saison**.

On va ensuite initiliser un driver pour effectuer notre web scraping à partir de la requête, précedemment créé. À noter que l'on écartera les compétitions suivantes : Coupe de France, Trophée des Champions, Coupe de France Féminine, afin de se focaliser sur les compétitions n'étant pas à élimination directe, et s'établissant plus sur le long-terme. Dans une volonté d'analyser la tendance récente, on ne s'occupera uniquement des 4 dernières saisons française (de 2021/2022 à 2024/2025).

Pour résumer les principales étapes du stockage de ces informations, on va logiquement dans un 1er temps se connecter au lien de la compétition présent dans

la table associé. On va ensuite fermé la page de cookie, pouvant bloquer la collecte de données. Pour information, sur ce site existe un menu déroulant afin de selectionner la saison de notre choix. On s'en servira (dans la mesure du possible) pour cliquer sur les saisons suivantes :

- **2021/2022**
- **2022/2023**
- **2023/2024**
- **2024/2025 (ou 2024/25)**.

Les informations essentielles figurant en 1ere page de ces saisons, on les collectera dans un objet, avant de passer à la saison suivante en effectuant la même opération. Les informations seront ensuite mis sous le format dataframe, et stocker dans la table season via la fonction insert_seasons. On fermera le driver une fois la tâche effectué.

Recherche des informations de chaque match et des équipes associés

Pour cette étape, on va récolter les informations de chaque match, et des équipes impliquées à partir des saisons stockées précedemment. On va ré-utiliser la même logique qu'auparavant :

- **Accession à la table season stockée sur Supabase**
- **Récupération des données à partir de la requête suivante : SELECT id_season, link_url FROM season, donnant l'identifiant de la saison et son lien url**
- **Création d'une classe Team, contenant l'identifiant de l'équipe, ainsi que son nom**
- **Création d'une fonction insert_team**
- **Création d'une classe Match, contenant l'indentifiant du match, saison, équipe à domicile et extérieur, la date du match, et enfin son lien url**
- **Création d'une fonction insert_matchs**

Étant donné la longueur plus conséquente de cette étape dans le but final de collecter les données sur les équipes et le match associé, on divisera chaque tâches par fonction. À noter que la logique reste semblable aux sections précédentes :

- **Ouvrir un driver pour notre web scraping**
- **Fermeture de la page de cookies lorsque cela est nécessaire**
- **Récupération de la page d'une saison à partir de la requête créé précedemment**
- **Stockage des informations suivantes pour chaque match de la journée en cours dans un object (en ne prenant pas en compte les matchs reportés, abandonné, ou donnant lieu à un tapis vert**
- **Appuyer sur la touche permettant d'accéder à la journée précédente lorsque tous les matchs de la journée courante ont été collecté**

- **Non prise en compte des saisons passées déjà collecté dans la recherche des données à insérer (permet d'accélerer la collecte des données)**
- **Stockage de nouveau des informations de ces matchs, et équipes, jusqu'à la 1ère journée, pour enfin passer à la saison suivante**
- **Mise des données au format dataframe**
- **Utilisation de la fontion insert_teams et insert_matchs pour stocker tout cela sur notre projet Supabase**
- **Fermeture du driver une fois toutes les tâches effectuées.**

Récupèration des informations sur les buts de chaque match

Enfin, pour cette étape, on va récolter les informations de but de chaque match stockées précedemment. On va ré-utiliser la même logique qu'auparavant :

- **Accession à la table info_match stockée sur Supabase**
- **Récupération des données à partir de la requête suivante : SELECT id_match, link_url, id_season FROM info_match, donnant l'identifiant du match, de sa saison et son lien url**
- **Récupération de l'information sur tous les identifiants des saisons et leur nom à partir de la requête suivante : SELECT DISTINCT s.id_season, s.season_name FROM Season s JOIN info_match im ON s.id_season = im.id_season; Cela permettra de demander à l'utilisateur les saisons qu'il souhaite stocker**
- **Récupération des identifiants de matchs déjà présent dans la base de données de buts, cela permettrant de ne pas récupérer la même information une nouvelle fois lorsque qu'elle a déjà été stocké précedemment**
- **Création d'une classe Goal, contenant l'identifiant du match, le score de l'équipe à domicile et à l'extérieur, le résultat du match, le nombre de buts inscrit par chaque équipe par tranche de 15 minutes et l'influence du 1er but sur le match**
- **Création d'une fonction insert_goals**

Encore une fois, étant donné la longueur plus conséquente de cette étape dans le but final de collecter les données sur les buts pour chacun des matchs, on divisera chaque tâches par fonction. Il est important de souligner que l'on demande à l'utilisateur les saison qu'il souhaite collecte dans une volonté réduire les chances que cela plante (dans le cas où trop de données sont à stocker). À noter que la logique reste semblable aux sections précédentes :

- **Initialisation d'un driver pour notre web scraping. À noter que les visualisations superflues seront enlevées afin d'accélerer la collecte des informations**
- **Fermeture de la page de cookies lorsque cela est nécessaire**
- **Récupération de la page d'une match à partir de la requête créé précedemment (sans prendre en compte les matchs déjà collectés)**

- **Accession à la section incidents contenant les faits marquants du matchs, dont sur chaque but du match et le score final**
- **Extraction du score du match via les colonnes homeScore et awayScore**
- **Déduction du résultat du match en fontion de la fonction précédente (Victoire à domicile, Nul ou Victoire à l'exterieur)**
- **Extraction des informations de buts par intervalles grâce aux données incidents.Toutes les colonnes seront égals à 0 si le score est nulle et vierge**
- **Déduction de l'influence du 1er but sur le match à partir de la fonction de result, et de l'information sur l'influence du 1er but**
- **Mise de toutes ces données au format dataframe**
- - **Non prise en compte des saisons passées déjà collecté, et des matchs déjà collectédans la recherche des données à insérer (permet d'accélerer la collecte des données)**
- **Réinitialisation du driver tous les 10 matchs, afin de réduire les chances que le code plante**
- **Utilisation de la fontion insert_goals pour stocker tout cela sur notre projet Supabase**
- **Fermeture du driver une fois toutes les tâches effectuées.**

Liste des idées de requêtes avant la mise en place de l'applications

La liste des variables que l'on va analyser :

- **l'influence du 1er but sur le match**
- **la distribution temporelles des buts**
- **l'influence du paramètre domicile/exterieur**
- **l'avantage du terrain x 1er but**
- **le nombre de buts par match**
- **Score du match**

Idées de requêtes réalisés:

- **Information sur le 1er but inscrit**

Information sur les proportions du 1er but inscrit au global

Information sur les proportions du 1er but inscrit pour chaque saison

Information sur les proportions du 1er but inscrit pour chaque compétition

Information sur les proportions du 1er but (inscrit ou encaissé) pour chaque équipe

Information sur les proportions du 1er but (inscrit ou encaissé) pour une équipe sur chaque saison

Information sur les proportions du 1er but inscrit de façon croissante pour une saison donné

- **l'influence du 1er but sur le match**

Information sur les proportions du 1er but (inscrit ou concédé) au global

Information sur les proportions de l'influence du 1er but inscrit pour chaque saison

Information sur les proportions de l'influence du 1er but inscrit pour chaque compétition

Information sur les proportions de l'influence du 1er but (inscrit ou encaissé) pour chaque équipe

Information sur les proportions de l'influence du 1er but (inscrit ou encaissé) pour une équipe sur chaque saison

Information sur les proportions de l'influence du 1er but (inscrit ou encaissé) de façon croissante pour une saison donné

- **la distribution temporelles des buts**

Proportion de buts inscrit par période au global

Proportion de buts inscrit par période de 15 min au global

Proportion de buts inscrit par période pour chaque saison

Proportion de buts inscrit par période de 15 min pour chaque saison

Proportion de buts inscrit par période pour chaque compétition

Proportion de buts inscrit par période de 15 min pour chaque compétition

Proportion de buts inscrit par période de 15 min pour une équipe

Proportion de buts encaissé par période de 15 min pour une équipe

Proportion de buts (inscrit ou encaissé) par période pour une équipe sur une saison donné

Proportion de buts (inscrit ou encaissé) par période pour une saison donné

Proportion de buts inscrit par période de 15 min (inscrit ou encaissé) pour une saison donné

- **l'influence du paramètre domicile/exterieur**

Proportion de victoire/nul/défaite au global

Proportion de victoire/nul/défaite pour chaque compétition

Proportion de victoire/nul/défaite pour chaque saison

Proportion de victoire/nul/défaite pour une équipe sur une saison donnée

Proportion de victoire/nul/défaite pour une saison donné pour chaque équipe

- **Statistique sur les buts**

Moyenne de but par match au global

Fréquence du score au global

Moyenne par but par saison

Fréquence du score pour une compétition donnée

Fréquence du score pour une saison donnée

Comparaison de la fréquence du score 1-1 selon les championnats

Comparaison de la fréquence du score 1-1 selon les compétitions

Moyenne de but par match pour une compétition

Moyenne de but par match pour une équipe sur une saison donnée

Fréquence des résultats (victoires, nuls, défaites) pour chaque équipe

Moyenne de buts inscrits et encaissés par équipe (domicile/extérieur) pour une saison donnée

In [ ]: