Projet Machine Learning

L’apnée est un sport de plus en plus populaire. De nombreuses compétitions existent autour du monde.
L’apnée, quels sont les facteurs qui peuvent influencer une plongée (en compétition)? Pour quelles raisons un apnéiste ne valide pas sa plongée?
C’est ce que vous allez devoir essayer de répondre dans ce projet.

https://www.youtube.com/live/8v_srQKy2_c?si=FVBlrPrXQYGiDcvl

Les résultats de ces compétitions venant de l’organisation d’AIDA peuvent être extrait duvsite web de l'organisation (https://www.aidainternational.org/Events/EventResults-3765)

Vous trouverez les données collectées en 07/23 sous format csv, focalisées sur les compétitions en eaux libres (depth competition).
Le fichier contient 26 842 plongées de 1984 à aujourd'hui.
La description des colonnes est celle-ci :

- Start : l’indice de départ de l’apnéiste dans la journée de compétition (peu intéressant)
- Diver : le nom de l’apnéiste et le trigramme de son pays
- Gender : le genre de l’apnéiste H/F
- Discipline : la discipline concurrue (FIM: immersion libre, CNF: brasse, CWT: poids constant libre souvent en monopalme, CWT-B: en by-palmes)
- Line : ligne de départ, très peu renseignée (peu intéressant)
- Official Top : l’heure officielle de son départ (peu intéressant?)
- AP : la profondeur annoncée, l’apnéiste doit annoncer la profondeur qu’il veut effecter et l’effectuer pour valider un carton blanc
- RP : la profondeur réalisée
- Card : le carton du juge, un blanc, la profonceur a été atteinte et le protocole de sortie réalisé, jaune, la profondeur n’a pas été atteinte ou le tag n’a pas été récupéré, le protocole de sortie bien effectuée, rouge: erreur sur le protocole de sortie
- Points : différence entre AP et RP et des points de pénalité le cas échéant
- Remarks : commentaire de la plongée
- Title Event : titre de la compétition
- Event Type : Type de compétition, profondeur ou piscine. Les données sont déjà filtrées sur le type == Depth competition ou Competition
- Day : jour de la compétition

Exploration des données (EDA)

Des erreurs de log , de collecte ou d’extraction de données peuvent être présentes (exemple le diver ()). L’exploration des données va vous permettre d’analyser ces données. Dans le

cas d'incohérence, des choix de transformation, suppression, remplacement peuvent être fait et ils doivent être justifiés.

Pour mieux comprendre les données, dans un premier temps, créer des graphiques permettant par la méthode de votre choix permettant de visualiser les données et de les comprendre, par exemple:

- la distribution du nombre de plongées par année
- la distribution du nombre d'apnéistes par année / par genre
- la distribution du nombre de cartons blancs / jaunes / rouges par discipline et par année
- toute autre visualisation permettant de comprendre la corrélation des données

Clustering

On souhaite expliquer pour quelles raisons un apnéiste obtient un carton rouge (la discipline, la profondeur annoncée, le manque d'expérience ?). Un carton rouge signifie que l'apnéiste n'a pas respecté le protocole de plongée ou de sortie qui assure que sa plongée s'est bien effectuée (la description est dans la colonne 'remark' mais elle est très peu normalisée). Pour cela, on va vouloir identifier les clusters qui permettront de définir le profil de la plongée échouée en utilisant les différentes caractéristiques obtenues par l'historique des plongées effectuées dans les compétitions.

1. Transformer les données catégorie en numérique (le clustering ne fonctionne pas sur le type de données catégorie, il faut les transformer):
    - la discipline (CNF, FIM, CWT, CWT-B, si autre à supprimer)
    - le genre (H, F, si autre à supprimer)
2. Créer comme caractéristique à partir des données :
    - le mois (ce qui donnera une indication sur la période de l'année où c'est déroulé la plongée)
    - le nombre de plongées effectuées avant la plongée (cumuler le nombre de plongée en regroupant par athlète triant par la date) == experience dive
    - le nombre de plongées effectuées avant la plongée (cumuler le nombre de plongée en regroupant par athlète et par discipline triant par la date) == experience discipline
    - (bonus) libre de créer d'autres caractéristiques numériques

	Diver	Day	Discipline	Month	Experience dive	Experience discipline
16592	Abdelatif Alouach (FRA)	2019-09-09	CNF	9	1	1
16633	Abdelatif Alouach (FRA)	2019-09-11	FIM	9	2	1
16728	Abdelatif Alouach (FRA)	2019-09-13	CWT	9	3	1
17939	Abdelatif Alouach (FRA)	2019-06-22	CNF	6	4	2
17968	Abdelatif Alouach (FRA)	2019-06-23	CWT	6	5	2
20072	Abdelatif Alouach (FRA)	2020-11-09	CWTB	11	6	1
20340	Abdelatif Alouach (FRA)	2021-05-12	CWTB	5	7	2
20386	Abdelatif Alouach (FRA)	2021-05-15	CWTB	5	8	3
20815	Abdelatif Alouach (FRA)	2021-06-19	CNF	6	9	3

Exemple des nouvelles colonnes (month, experience dive, experience discipline) sur les plongées qui correspondent à l'apnéiste 'Abdelatif Alouach (FRA)'

3. Sélectionner les colonnes pertinentes au problèmes (indice: les numériques, et celles connues avant une plongée)
4. Normaliser les données pour que toutes les colonnes soient comparables en terme de distance euclidienne
5. Appliquer **deux méthodes de clustering (K-means et DBSCAN) pour détecter si plusieurs profils de plongée se dessine** :
    a. sur les plongées qui ont obtenu un carton Rouge (card==RED)
    b. sur les plongées qui ont obtenu un carton Blanc (card==WHITE) (pour comprendre l'effet inverse, lorsque les plongées se passent bien)
6. Décrire et visualiser les résultats obtenus (vous pouvez afficher selon 2 ou 3 caractéristiques). Quels sont les différents clusters trouvés et leurs caractéristiques?
7. Expliquer comment vous avez choisi les hyperparamètres de chacune des méthodes
    a. le nombre de clusters pour K-means. Utiliser la méthode du coude, Elbow method)
    b. La densité pour DBSCAN
8. Commenter les résultats obtenus, pour et contre chaque méthode. Est-ce que les méthodes de clustering sont adaptées au problème?

Classification

1. Prédire si le résultat d'une plongée va être un carton blanc, jaune, rouge selon les caractéristiques qui paraissent pertinentes au problème. Utiliser le classifieur SVM avec **différents noyaux**.
2. Justifier vos choix (en phrase et par de la visualisation si nécessaire).
3. Créer un tableau récapitulant vos différents résultats.
4. Commenter les résultats obtenus.
5. Explorer d'autres pistes de classifieurs (au moins un) et comparer les résultats

BONUS

Les syncopes (malaise de l'apnéiste) sont identifiés par la colonne 'Remark' normalement par le commentaire 'DQBO', un critère plus précis que le carton pour identifier une plongée qui ne s'est pas déroulée comme prévue. Mais cette colonne n'est pas normalisée.
En bonus, vous pouvez regrouper essayer de la normaliser et d'appliquer une classification sur cette colonne (syncope ou non)

Evaluation

Pour le 07/12/23, pour la partie rendu écrite :

- Il est demandé le code source pour chaque partie.
- Le code doit être lisible, clair et structuré. La documentation est la clé.
- Des graphiques et leur explications sont demandés dans toutes les parties. Un graphique sans description ne sera pas pris en compte.

Un oral de 10 min sera demandé avec :

- Explication des méthodes utilisées pour la partie clustering , la justification, les pour et les contre
- Description des résultats de classification et les caractéristiques utilisées.

Group 1	Group 2
1 Quentin 2 Gianluca 3 Darius	1 Mohammed 2 Léo 3 Maxime
Group 3	**Group 4**
1 Benoît 2 Amin 3 Thomas	1 Thibaut 2 Mohammad 3 Maïa
Group 5	
1 Sacha 2 Alexis 3 Luna	