Projet Machine Learning

L'apnée est un sport de plus en plus populaire. De nombreuses compétitions existent autour du monde.
Une plongée réussie en compétition est la réussite de la profondeur annoncée. Des juges sont présents et donnent un carton selon la réussite de la plongée :

- blanc : l'apnéiste a été à la profondeur annoncée et a effectué le protocole de sortie
- jaune: l'apnéiste a tourné avant (la profondeur annoncée n'est pas atteinte) et a effectué son protocole de sortie
- rouge: l'apnéiste n'a pas effectué le protocole de sortie (syncope, défaut de protocole, discipline non respectée). La plongée n'est pas validée.

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/W6Hkdo4h2W0?si=CMy-Y67CVPQZaa9k

Les résultats des compétitions venant de l'organisation d'AIDA peuvent être extrait du site web de l'organisation (https://www.aidainternational.org/Events/EventCalendar)

Vous trouverez les données collectées en 11/2024 sous format csv, focalisées sur les compétitions en eaux libres (Depth Competition).

Le fichier contient le résultat des plongées de 1984 à aujourd'hui.
La description des colonnes est celle-ci. Les colonnes suivantes correspondent aux caractéristiques :

- Start : l'indice de départ de l'apnéiste dans la journée de compétition (peu intéressant)
- Diver : le nom de l'apnéiste
- **Nationality** : le pays de l'apnéiste
- **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
- **Day** : jour de la compétition
- Title Event : titre de la compétition
- Event Type : le type de l'événement

Les informations obtenues après la plongée (ce ne sont donc pas des caractéristiques mais la sortie):

- *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

Exploration des données (EDA)

Des erreurs de log , de collecte ou d'extraction de données peuvent être présentes. 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.

1. Explorer/ décrire en statistique chaque colonne, vérifier les données manquantes ou les données incohérentes. Pour chaque colonne, indiquer quelle stratégie vous adoptez pour les données incohérentes ou manquantes (suppression, remplacement …). Supprimer les lignes correspondantes si nécessaire. Justifier.
2. Convertir les colonnes qui doivent en numérique (AP et RP)

3. Pour mieux comprendre les données, créer des graphiques permettant par la méthode de votre choix permettant de visualiser les données et de les comprendre
    a. la distribution du nombre de plongées par année
    b. la distribution du nombre d'apnéistes par année / par genre / par nationalité
    c. la distribution du nombre de cartons blancs / jaunes / rouges par discipline et par année
    d. 2 autres graphiques libres expliquant les données

    Décrire chaque graphique.

4. Créer comme caractéristiques à partir des données :
    - **month** = le mois (ce qui donnera une indication sur la période de l'année plutôt que le jour précis)
    - **experience_dive** = le nombre de plongées effectuées avant la plongée (cumuler le nombre de plongée en regroupant par athlète triant par la date)
    - **experience_discipline** = le nombre de plongées effectuées avant la plongée (cumuler le nombre de plongée en regroupant par athlète et par discipline triant par la date)

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

Diver	Day	Discipline	Month	Experience dive	Experience discipline
Abdelatif Alouach (FRA)	2019-09-09	CNF	9	1	1
Abdelatif Alouach (FRA)	2019-09-11	FIM	9	2	1
Abdelatif Alouach (FRA)	2019-09-13	CWT	9	3	1
Abdelatif Alouach (FRA)	2019-06-22	CNF	6	4	2
Abdelatif Alouach (FRA)	2019-06-23	CWT	6	5	2
Abdelatif Alouach (FRA)	2020-11-09	CWTB	11	6	1
Abdelatif Alouach (FRA)	2021-05-12	CWTB	5	7	2

5. Créer des graphiques par rapport aux nouvelles caractéristiques calculées par rapport à la couleur du carton. Apportent t-elles des informations ?

6. Création du dataframe pour les apprentissages et qui va servir pour la suite des questions:
    a. **Sélectionner les colonnes caractéristiques pertinentes pour la suite (.**
    b. **Créer une colonne 'target' qui correspond à la colonne 'Card'**

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, CWTB, si autre à supprimer)
    - la nationalité
    - le genre (H, F, si autre à supprimer)

2. Normaliser les données pour que toutes les colonnes soient comparables en terme de distance euclidienne

3. Appliquer **deux méthodes de clustering (K-means et DBSCAN) pour détecter si plusieurs profils de plongée se dessinent** :
    a. sur les plongées qui ont obtenu un carton Rouge (card==RED), créer un dataframe ne comportant que ces plongées et effectuer le clustering. Attention à ne pas prendre comme caractéristique la colonne 'card'.
    b. sur les plongées qui ont obtenu un carton Blanc (card==WHITE) (pour comprendre l'effet inverse, lorsque les plongées se passent bien), créer un dataframe ne comportant que ces plongées et effectuer le clustering. Attention à ne pas prendre comme caractéristique la colonne 'card'.

4. Décrire et visualiser les résultats obtenus (vous pouvez afficher selon 2 ou 3 caractéristiques). Quelles sont les distances inter et intra clusters? Est-ce possible d'avoir des profils 'type' selon le carton obtenu?

5. 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

6. 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. 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 sur différentes métriques **(matrices de confusion, accuracy, précision, rappel).**
4. Commenter les résultats obtenus.
5. Explorer d'autres pistes de classifieurs (**au moins deux autres**) et comparer les résultats avec le meilleur SVM obtenu.

Evaluation

Pour le 26/11/2024, 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.

Une présentation de 10 min sera demandée avec :

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