Meddas Kilian, Sriwelavan Theeban

**Documentation du Projet de Machine Learning**

**1. Introduction**

Ce projet de Machine Learning explore et prédit des événements liés aux performances en plongée, avec un accent particulier sur les types de cartons attribués aux plongeurs (blancs, jaunes, rouges). L'objectif principal est de révéler des corrélations pertinentes, de segmenter les participants, et de développer des modèles de prédiction robustes.

**Objectifs du projet**

1. Comprendre et analyser le dataset disponible.
2. Découvrir des relations clés entre les variables, comme l'expérience et les cartons reçus.
3. Segmenter les plongeurs à l’aide de méthodes de clustering.
4. Construire et évaluer des modèles prédictifs pour la classification des cartons.
5. Documenter les conclusions et expliquer les résultats de manière détaillée.

**2. Dataset**

**Structure des données**

Le dataset contient plusieurs variables importantes :

* **Announced Depth (AP)** : Profondeur annoncée par le plongeur.
* **Reached Depth (RP)** : Profondeur effectivement atteinte.
* **Points** : Score obtenu pour la performance.
* **Discipline** : Type de plongée :
  + CNF : Constant Weight No Fins.
  + CWT : Constant Weight.
  + CWTB : Constant Weight Bi-Fins.
  + FIM : Free Immersion.
* **Nationality** : Nationalité des participants.
* **Gender** : Sexe des plongeurs.
* **Experience (Dive et Discipline)** : Mesure de l’expérience cumulée.
* **Carton (Card)** : Carton reçu (blanc, jaune ou rouge).

**3. Analyse exploratoire des données (EDA)**

**Visualisations principales**

1. **Répartition des disciplines**
   * Les disciplines **CWTB** (36,7%) et **FIM** (30,1%) dominent dans le dataset, tandis que **CNF** est la moins pratiquée (18%).
2. **Évolution des points moyens au fil des années**
   * Les données montrent une baisse globale des points entre 2018 et 2020, suivie d’une amélioration jusqu’en 2022, avant une nouvelle diminution en 2023.
3. **Distribution des cartons par couleur et discipline**
   * Les proportions de cartons (blancs, jaunes, rouges) varient selon chaque discipline, offrant des perspectives sur les performances.
4. **Relation entre l’expérience et les points obtenus**
   * Les plongeurs plus expérimentés tendent à recevoir davantage de cartons blancs et à obtenir de meilleurs scores.

**4. Clustering**

**Méthodologie**

Pour regrouper les plongeurs selon leurs caractéristiques, nous avons utilisé l’algorithme **K-Means**.

**Méthode du coude**

Le graphique indique un "coude" significatif à **3 ou 4 clusters**, suggérant ces options comme optimales.

**Résultats du clustering**

Les plongeurs sont segmentés selon leur expérience et leurs performances (AP et RP), permettant de distinguer des groupes aux compétences et comportements variés :

* Cluster de plongeurs novices avec moins de réussite.
* Cluster de plongeurs intermédiaires avec des performances stables.
* Cluster d’experts ayant des scores élevés et des échecs réduits.

**5. Modèles prédictifs**

**Objectif**

L’objectif est de prédire le type de carton (blanc, jaune ou rouge) qu’un plongeur pourrait recevoir.

**Algorithmes testés**

1. **SVM (Support Vector Machines)**
   * Différents noyaux ont été testés : linéaire, polynomial, RBF. Le noyau linéaire a donné les meilleures performances.
2. **Random Forest**
   * Bonnes performances générales, bien qu’il y ait des confusions pour les cartons rouges.
3. **KNN (K-Nearest Neighbors)**
   * Résultats moins précis, avec de nombreuses confusions entre cartons jaunes et rouges.

**Métriques d’évaluation**

* **Matrice de confusion** : Visualise les erreurs de classification.
* **Accuracy, Precision, Recall, F1-Score** : Mesures utilisées pour évaluer chaque modèle.

**6. Résultats clés**

**Relations significatives observées**

1. Les plongeurs expérimentés sont plus susceptibles de recevoir des cartons blancs.
2. Les disciplines influencent les scores obtenus :
   * Les disciplines **FIM** et **CWTB** sont associées à de meilleures performances.
3. Les cartons rouges sont souvent liés à des erreurs ou à une expérience limitée.

**Meilleur modèle**

Le modèle **Random Forest** a montré le meilleur équilibre entre précision et rappel pour prédire les cartons.

**7. Conclusion**

**Apports du projet**

* Une meilleure compréhension des facteurs influençant les performances et la réception des cartons.
* Une segmentation efficace des participants en fonction de leur expérience.
* Une prédiction fiable des cartons, utile pour les analyses futures et la préparation des compétitions.

**Prochaines étapes**

1. Affiner les modèles en explorant d’autres algorithmes.
2. Approfondir l’analyse des disciplines spécifiques, comme les variations régionales.
3. Exploiter les clusters pour des études qualitatives supplémentaires.

**8. Annexes**

**Documentation des variables**

| **Variable** | **Description** |
| --- | --- |
| **AP** | Profondeur annoncée par le plongeur |
| **RP** | Profondeur atteinte par le plongeur |
| **Points** | Points obtenus |
| **Carton** | Carton reçu (blanc, jaune, rouge) |
| **Discipline** | Type de plongée (CNF, CWT, CWTB, FIM) |
| **Experience\_Dive** | Expérience totale en plongée |
| **Experience\_Discipline** | Expérience par discipline |

**Acronymes des disciplines**

* **CNF** : Constant Weight No Fins.
* **CWT** : Constant Weight.
* **CWTB** : Constant Weight Bi-Fins.
* **FIM** : Free Immersion.