Analyses et résultats des données de l'égypte : Le fichier Egypte_period_prediction.csv

chahrazed labba

le 17 Mai 2023

1 Prédire la période temporelle avec 17 étiquettes

Cette section décrit les données utilisées et présente les résultats des prédictions obtenues en utilisant les algorithmes de machine learning suivants : RF, svm et xGboost.

1.1 Les données

- Nombre d'individus : 315
- Les caractéristiques (features) : 'classe_age','Position_corps','Orientation_tete','Orientation_Face', 'Position_mains','Nb_actes','Mobilier','Nb_Mobilier','mob_tete','mob_membre_inf','mob_thorax', 'mob_dos','mob_devant_corps','mob_contre_jarre','mob_dans_jarre','mob_exterieur_coffre', 'mob_dans_coffre','mob_contre_paroi', 'mob_dans_remplissage','Signalisation','pelle ','Ornement ', 'Nb_Ornement','orn_Porte','orn_Nonporte', 'Coquillage','Aspatharia_unio', 'Perles','Fard','Outils','Prosterne','Decoupe'

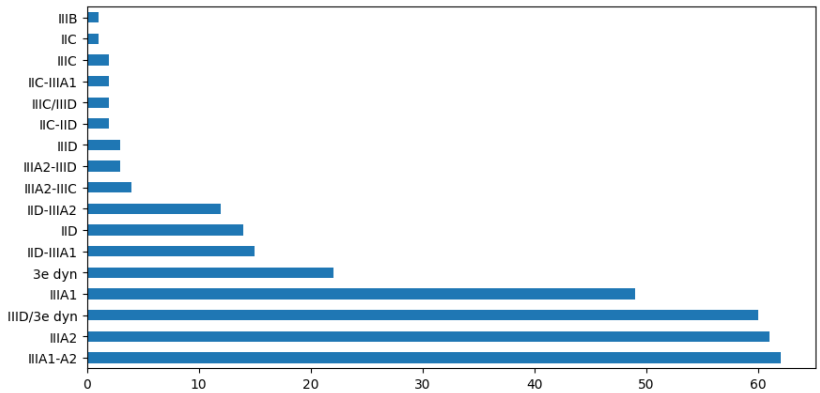

Figure 1: Nombre des individus par class label

La répartition des données entre les différentes étiquettes de classe présente un énorme déséquilibre. Certaines étiquettes de classe ne contiennent qu'un seul échantillon.

	A totale	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16
RF (%)	37.14	75.0	53.22	42.62	22.44	9.09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM (%)	33.01	58.33	45.16	29.50	20.40	27.27	6.66	21.42	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Xgboost(%)	35.87	66.66	43.54	36.06	32.65	22.72	0.0	21.42	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 1: Accuracy globale et accuracy par class label

les étiquettes de classe et leur signification

- IIID/3e dyn:0
- 'IIIA1-A2':1
- 'IIIA2':2
- 'IIIA1':3
- 3e dyn':4
- IID-IIIA1':5
- 'IID':6
- 'IID-IIIA2':7
- IIIA2-IIIC':8
- 'IIID':9
- 'IIIA2-IIID':10
- 'IIIC':11
- 'IIIC/IIID':12
- 'IIC-IID':13
- IIC-IIIA1':14
- 'IIC':15
- 'IIIB':16

Pour les trois modèles, les étiquettes de classe de 8 à 16 n'ont jamais été prédites. Ce problème est dû à un problème de représentation des données (sous-représentation par rapport au reste des étiquettes de classe).

- df_results_SVM_17_labels_LOO.csv (contient les prédictions faites par SVM)
- df_results_RF_17_labels_LOO.csv (contient les prédictions faites par RF)
- df_results_xgboost_17_labels.csv (contient les prédictions faites par Xgboost)

2 Prédire la période temporelle avec 5 étiquettes

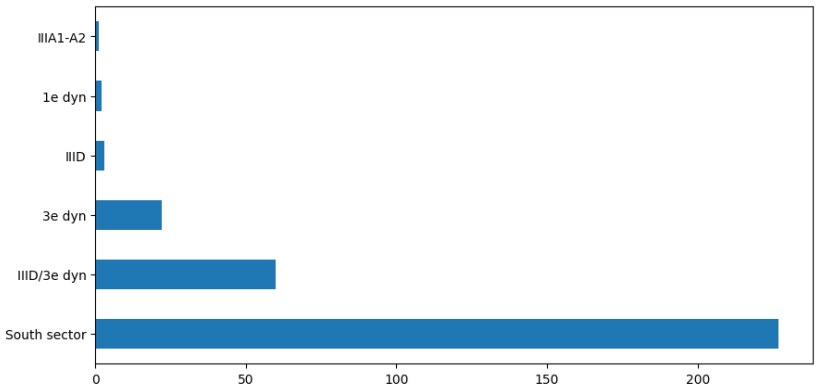

Figure 2: Nombre des individus par class label

La répartition des données entre les différentes étiquettes de classe présente un énorme déséquilibre. Certaines étiquettes de classe ne contiennent qu'un seul échantillon.

- 'South sector':0
- 'IIID/3e dyn':1
- '3e dyn':2
- 'IIID':3
- '1e dyn':4
- 'IIIA1-A2' : 5

	Accuracy	0	1	2	3	4	5
RF	78.41	99.55	35.0	0.0	0.0	0.0	0.0
SVM	77.46	92.95	45.0	27.27	0.0	0.0	0.0
Xgboost	78.41	94.27	46.66	22.72	0.0	0.0	0.0

Table 2: Accuracy globale et accuracy par class label

3 Résultat de clustering en utilisant k-means

Le clustering donne lieu à deux clusters.

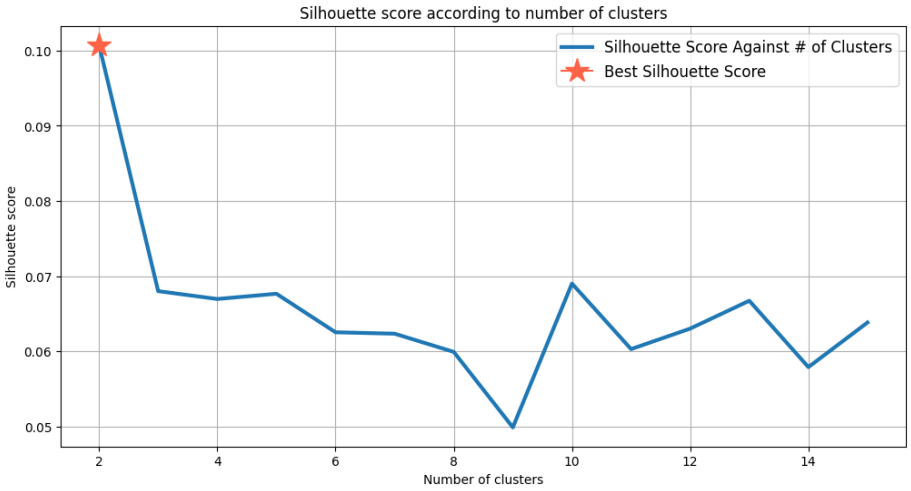

Figure 3: Nombre de clusters

Comme K-means est une boîte noire et que nous ne pouvons pas savoir quelles caractéristiques ont été utilisées pour réaliser le clustering, j'ai appliqué un algorithme de classification (Random Forest), puis la lib Shap pour expliquer le clustering.

Le fichier Egypt_Cluster_Data_315_V.csv contient les résultats de clustering.

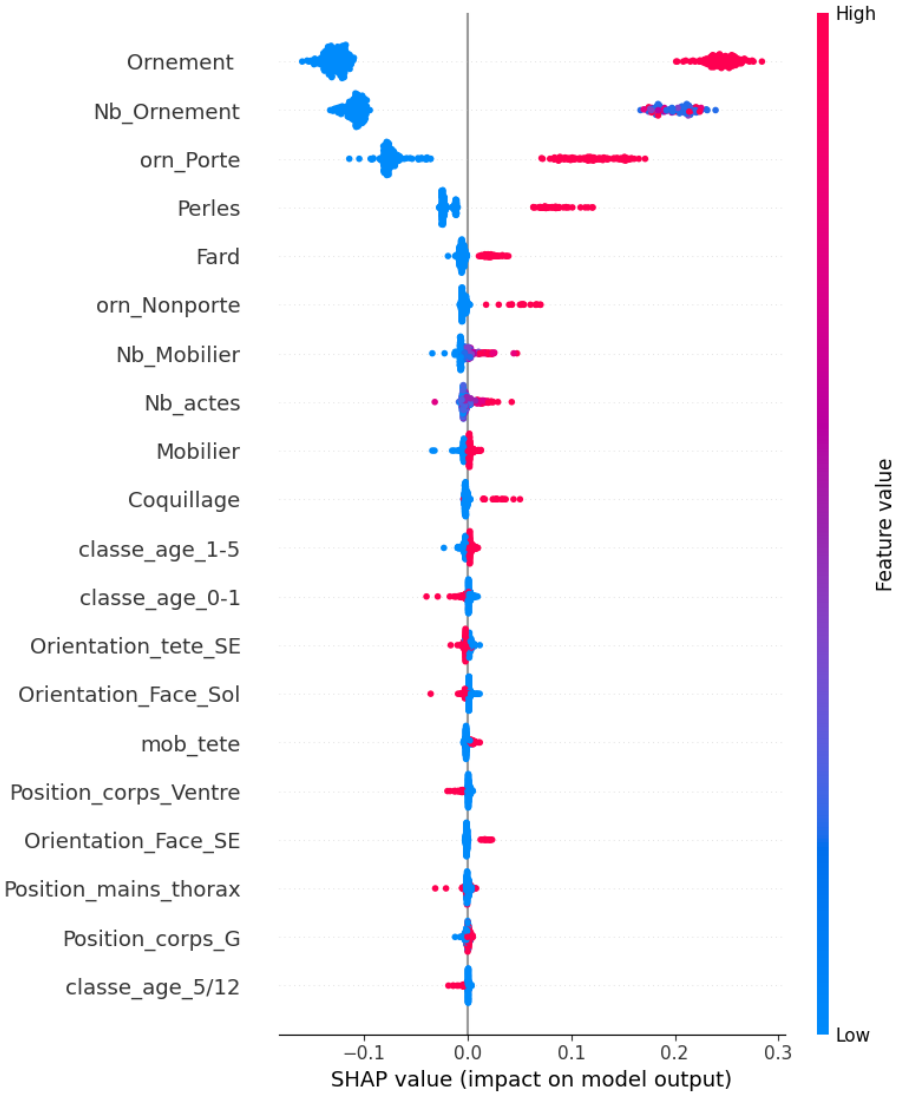

Figure 4: Cluster 0

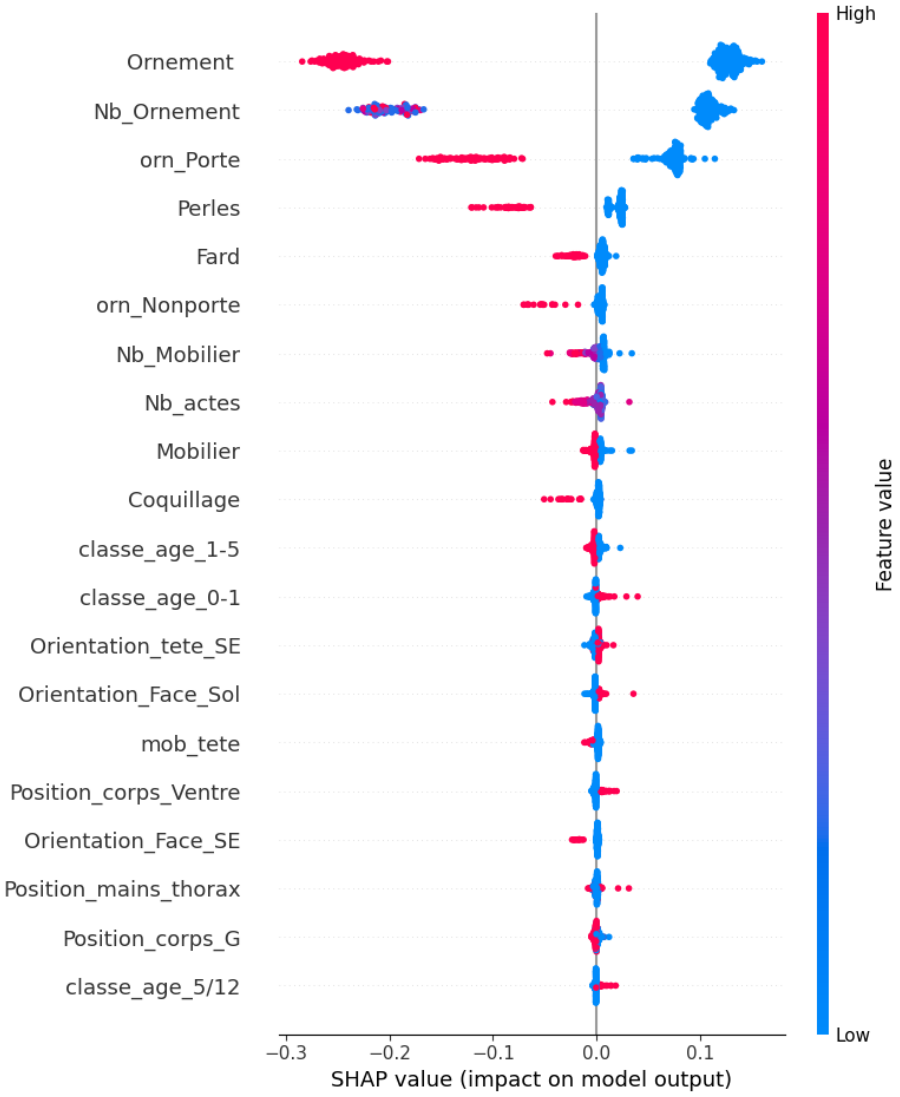

Figure 5: Cluster 1