## Summary (/projepts/jepts

## ML - Machine Learning

## Ce slide contient ce que nous avons pu faire pour la partie Machine Learning.

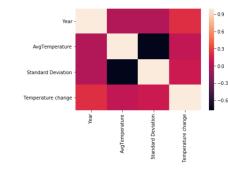
En regardant un peu nos datasets, nous n'avons pas trouvé, à part celui des sur les températures après iointure, de datasets pertinents pour faire de la régression ou de la classification,

Nous avons donc réalisé des **régressions** sur le dataset contenant les températures movennes, les variations et les déviations des températures

 $(/projects/PROJET\_BIUM/dashboards/insights/zB8DTj4\_temperaturejoined analysis-table/view)$ 

Temperature_joined_analysis table						
Country ()	Year ()	Month ()	AvgTemperature ()	Standard Deviation ()	Temperature c	
Algeria	1995	January	51.41935483870968	1.294		
Algeria	1995	February	54.7999999999999	1.3619999999999999		
Algeria	1995	March	55.0	1.188	0.00900000	
Algeria	1995	April	57.19333333333334	0.9420000000000001		
Algeria	1995	May	67.40967741935485	1.032		
Algeria	1995	June	71.72333333333333	0.731		
Algeria	1995	July	77.50000000000001	0.614		
Algeria	1995	August	77.95517241379311	0.55		
Algeria	1995	September	71.73	0.43799999999999994	1.019000	
Algeria	1995	October	68.52580645161292	0.794		
Algeria	1995	November	61.70333333333333	1.011		

Heatmap des corrélations entre les variables



Nous observons une corrélation de -0.769915 entre les variables Standard Deviation et AvgTemperature.

Standard Deviation et AvgTemperature sont assez grandement négativement corrélées

Lorsqu'une de ces deux variables augmente, l'autre diminue plus ou moins.

Nous allons essayé de prédire la température moyenne en fonction des autres variables et en prenant en compte l'ordre ologique de la colonne Year.

Year AvgTemperature Standard Deviation Temperature change

1.000000	0.008899	0.008518	0.225228
0.008899	1.000000	-0.769915	0.066880
0.008518	-0.769915	1.000000	0.114385
0.225228	0.066880	0.114385	1.000000
	0.008899	0.008899 1.000000 0.008518 -0.769915	0.008899 1.000000 -0.769915   0.008518 -0.769915 1.000000

ns ensuite testé différents modèles, hyper-paramètres, variables et paramètres d'évaluation pour essayer de trouver le modèle le plus pertinent

Pour cela, nous avons séparé les données initiales en deux parties, une partie d'apprentissage qui contient 80% des données, et une autre partie qui représente les données de test et qui contient les 20% des données restantes

 $Les mod\`eles seront alors entrain\'es sur les donn\'ees d'apprentissage puis \'evalu\'es sur les donn\'ees de test.$ 



Nous pouvons voir ci-dessous que pour les deux modèles, la variable qui sert vraiment pour prédire la température moyenne est la Standard Déviation.

Ceci n'est pas surprenant, cette observation est cohérente avec les **analyses de corrélations** précédemment effectuées.

XGBoost qui avait l'air plus performant, prend beaucoup en compte pour sa prédiction, les variables Temperature Change et Year par rapport au modèle Random Forest.

Nous pouvons également constater que le pays qui sert le plus est la Mongolie, et cela peut s'expliquer par le fait que c'est le pays avec la température moyenne la plus basse au monde dans nos données d'après nos analyses de la partie EDA - Température.

RANDOM\_FOREST\_REGRESSION XGBOOST\_REGRESSION