

Rapport du Projet Métier

Filière :

Génie Industriel : Data Science et Intelligence Artificielle

Sous le thème :

**Prédiction de la qualité des eaux du
fleuve Moulouya**

Réalisé par :



BAGUIAN Harouna



DIASSANA Fatoumata

Sous l'encadrement de :



Mr. Mansrour



Mr. Mansouri

Année Universitaire : 2021-2022

REMERCIEMENTS

Avant tout développement et rédaction de cette expérience, il apparaît opportun pour nous, de commencer par nos profonds remerciements à Mr. Tawfik Mansrour et Mr. Imad Mansouri, nos chers encadrants qui nous ont guidé et soutenus tout au long de la réalisation de ce projet.

Ensuite nous tenons à remercier également Mr MASROUR, le responsable de la filière Génie Industriel option Data Science et Intelligence Artificielle et Mr le Directeur Général de l'ENSAM qui nous ont offerts l'opportunité d'avoir cette formation industrielle et en Data Science et Intelligence Artificielle. Enfin, nous remercions également nos chers professeurs de la Filière Génie Industriel option Data Science et Intelligence Artificielle pour leurs guides à la réussite de ce projet.

RESUME

L'hygiène et la sécurité alimentaire est de nos jours une préoccupation mondiale, prédire la qualité des eaux constitue un socle pour répondre à cette problématique.

Plusieurs normes existent pour prédire la qualité des eaux de surfaces à l'instar de la norme marocaine. Cette norme utilise plusieurs indicateurs. Passer entre les lignes de ces différentes normes peut s'avérer pénible. C'est dans cette optique que nous voulons utiliser l'intelligence artificielle plus précisément la machine learning pour répondre à cette problématique.

A cours de ce projet nous avons développé une application pour prédire la qualité des eaux de surface accessible à tous, aux professionnelles et aux non initié à l'intelligence artificielle.

Table des matières

I.	Accueil	5
II.	Preprocessing.....	5
1.	Uploader les données	5
2.	Dataset description	6
2.a	First look.....	6
2.b	Resume.....	7
3.	Preprocessing.....	7
3.a	La suppression des missing values	8
3.b	Imputation statistique par la moyenne	8
3.c	Visualisation et suppression des outliers.....	9
4.	Features selection	10
5.	Compare.....	11
6.	La normalisation des données	12
III.	Data visualisation	12
1.	Scatter plot.....	12
2.	Box plot	13
3.	Corrélation matrix.....	13
4.	Count plot	13
5.	Distribution plot.....	14
IV.	Machine Learning.....	14
1.	La classification	15
1.a	SVC	16
i.	Définition	16
ii.	Paramétrage du modèle (parameters tuning)	16
iii.	Résultats.....	17
1.b	KNN	17
i.	Definition	17
ii.	Paramétrage du modèle (parameters tuning)	18
iii.	Résultat du modèle.....	18
1.c	RandomForest Classifier	19
i.	Définition	19
ii.	Paramétrage du modèle (parameters tuning)	19
iii.	Résultats obtenus	19

1.d	Tree	20
i.	Définition	20
ii.	Paramétrage du modèle (parameters tuning)	21
iii.	Resultats.....	21
1.e	Voting.....	22
	Définition	22
2.	Régression.....	22
V.	Prédiction.....	24
1.	Predire pour une seule mesure	25
2.	Predire en uploadant un fichier csv	26
VI.	Le rapport.....	27
VII.	Le map.....	30
	Conclusion	32

Introduction

Le machine Learning est en plein essor, de nos jours les algorithmes sont devenu de plus en plus efficace et donne de bon résultats. les algorithme de machine Learning il y'a de toute sorte allant de la prédiction a la classification. Cet essor est favorisé par le boum de la masse de donnée. Ces données doivent être traité afin de les utilisé dans l'algorithme de machine

Learning c'est le data preprocessing. Le traitement des données fait référence à la conversion de données brutes en informations significatives, et ces données sont également lisibles par machine. Ainsi, le traitement des données implique la suppression des valeurs aberrantes, les outliers et les normalisé les données. Après le traitement des données une visualisation est nécessaire afin de mieux comprendre nos données, c'est le data visualisation. La data visualisation (data viz ou représentation graphique de données ; elle s'écrit également data visualization) consiste à structurer visuellement des données recueillies et stockées. Ainsi, l'exploitation des données se fait plus facilement. Ensuite nous pouvons choisir le modèle adéquat pour entrainer avec nos données. Faire tout ceci est réservé a un public aguerri.

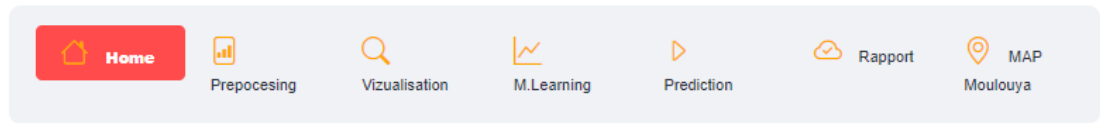
Mais si on pouvait faire du machine Learning sans coder ?

C'est au vu de répondre a cette problématique que nous avons créé data cleaner.

Mots clés : python, data preprocessing, data visualisation, streamlit, prédiction, classification

Le machine Learning passe par plusieurs étapes : data collection and preprocessing, choix du modèle, training du modèle, évaluation et la prédiction

I. Accueil



Machine Learning Platform

Bienvenue sur notre plateforme de data science
realisée par: BAGUIAN HAROUNA ET DIASSANA
FATOUMATA!

Suivez nous sur Github:

• [DIASSANA Fatoumata/GitHub](#)

II. Preprocessing

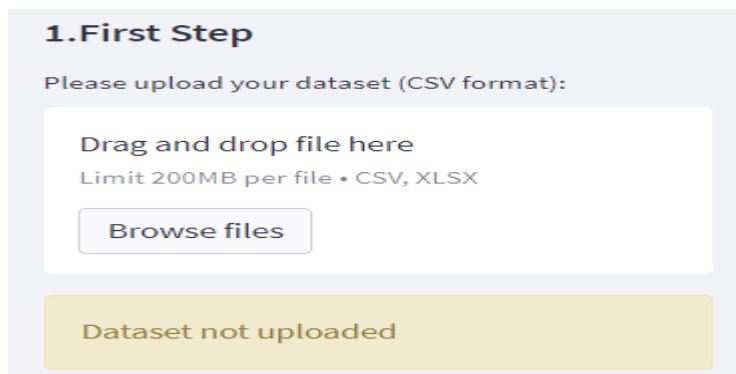
Le preprocessing est la deuxième étape du machine Learning juste après la collecte des données.

Le prétraitement fait référence aux transformations appliquées à nos données avant de les alimenter à l'algorithme. Le prétraitement des données est une technique utilisée pour convertir les données brutes en un ensemble de données propres.

Pour le prétraitement de nos données on est passé par plusieurs étapes.

1. Uploader les données

Dans cette partie les données brutes avec lesquels on désire entrainer les algorithmes sont uploader sous format csv.



1. First Step

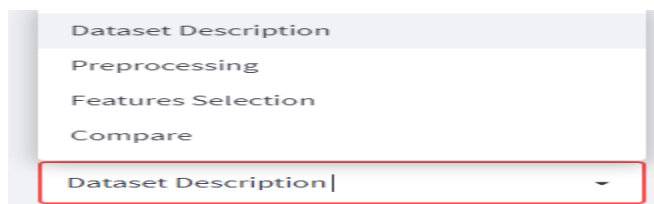
Please upload your dataset (CSV format):

Drag and drop file here
Limit 200MB per file • CSV, XLSX

[Browse files](#)

Dataset not uploaded

2. Dataset description



Dataset Description

Preprocessing

Features Selection

Compare

Dataset Description | ▾

Comprendre le dataset est très primordial pour faire du machine Learning.

Plusieurs informations très utiles peuvent émaner d'un dataset. Connaitre ses informations est très crucial. Parmi ces informations on peut citer le nombre de ligne et de colonne du dataset, le nombre de lignes, de colonnes du dataset, les différents types des variables (features), le nombre de valeurs manquantes (qui peuvent provoqué des erreurs d'entraînement pour des modèles de machine Learning qui sont sensible au valeurs aberrantes).

Après avoir uploadé le bon dataset, sélectionné dans le select box juste en bas du message « dataset uploaded successfully » l'option dataset description.

2.a First look

Le radio bouton first look nous donne un aperçu très global du dataset que vous venez t'uploader.

Ces informations sont le nombre et lignes et de colonnes, le nombre de variables catégorielle et numérique, le nombre de valeurs manquantes et leurs pourcentages, on peut également avoir les statistiques de base comme la moyenne, l'écart type, la valeur minimale et maximale de chaque variable.

1. First Step

Please upload your dataset (CSV format):

Drag and drop file here
Limit 200MB per file • CSV, XLSX

Browse files

Classeur11.csv
4.2KB

Dataset uploaded successfully!

2. Second Step

What do you want to explore?

Dataset Description

Choose

☒ First Look

☐ Resume

Exploring the dataset 🔍

	PH	T	CE	O2	NH	NO	SO	PO	DBOS	IQE	Classe
0	2.4080	0.3510	0.0020	4.9400	2.5060	0.0310	0.0240	0.2850	1.9930	17.0060	1
1	2.4890	0.3890	0.0020	4.3880	2.5060	0.0220	0.0320	0.3130	2.0500	16.1410	1
2	2.4640	0.4520	0.0020	4.1900	2.3920	0.0360	0.0290	0.3700	2.1530	15.8530	1
3	2.7110	0.5160	0.0020	4.2130	7.4040	0.0280	0.0310	0.4270	2.3350	21.3640	1
4	2.5100	0.7060	0.0030	3.6900	5.6950	0.0470	0.0730	42.9990	6.8910	65.0980	4

- The dataset has 66 observations and 11 variables. Hence, the *instances-features* ratio is ~6.
- The dataset has 0 categorical columns (e.g.) and 11 numerical columns (e.g. PH).
- Total number of missing values: 0 (~0.00%).

Description 🖨️

	PH	T	CE	O2	NH	NO	SO	PO	DBOS	IQE	C
count	66.0000	66.0000	66.0000	66.0000	66.0000	66.0000	66.0000	66.0000	66.0000	66.0000	66.0000
mean	2.6083	0.6435	0.0047	4.0682	41.1839	0.1660	0.0831	19.1663	6.6794	77.2518	1.0000
std	0.1495	0.1548	0.0024	0.8845	98.9711	0.2721	0.0620	23.4124	8.0559	122.4685	0.0000

2.b Resume

Comme le radio bouton first look, résumé donne en plus d'une description basique il donne une description très poussé et intégrale du dataset, on peut observer les différentes corrélations entres les variables. Naviguez jusqu'en bas de la page pour bien analyser votre dataset

Overview

Overview

Alerts 43

Reproduction

Dataset statistics

Number of variables	11
Number of observations	66
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	5.8 KiB
Average record size in memory	89.9 B

Variable types

Numeric	10
Categorical	1

Active Windows

3. Preprocessing


Comme définit plus haut le preprocessing ou le traitement des données fait référence aux transformations appliquées à nos données avant de les alimenter à nos algorithmes.

03 grandes notions sont abordées dans cette partie, la suppression des valeurs aberrantes (missing values), imputation statistique par la moyenne, la suppression des outliers.

3.a La suppression des missing values

Le problème de la valeur manquante est assez courant dans de nombreux ensembles de données réels. Une valeur manquante peut fausser les résultats des modèles d'apprentissage automatique et/ou réduire la précision du modèle.

La suppression se fait par colonne, nous définissons des seuils(de 50 à 100) c'est à dire le pourcentage maximal de valeurs manquante par colonne et nous supprimons la colonne si le pourcentage dépasse ce seuil.

Missing values 

Choose seuil NaN percentage

100|

100

90

80

70

60

50

3.b Imputation statistique par la moyenne

Après avoir supprimer les valeurs colonnes qui répondaient a notre critère de seuil il faut maintenant remplacer les valeurs manquante entre les lignes du dataset par la moyenne, ça s'appelle l'imputation statistique par la moyenne, il y'a plusieurs types d'imputation a l'instar de celle de la moyenne, on a l'imputation par la médiane (ce n'est pas une bonne solution car la médiane peut être une valeur manquante) il y'a aussi l'imputation par zéro (ces zéros risque d'être considéré comme des outliers)

Numeric features Imputation

--imputation--|

--imputation--

Mean

Mediane

Zero

3.c Visualisation et suppression des outliers

Plusieurs algorithmes de Machine Learning sont sensibles aux données d'entraînement ainsi qu'à leurs distributions. Avoir des *Outliers dans Training Set* d'un algorithme de Machine Learning peut rendre la phase d'entraînement plus longue. Sans mentionner que l'apprentissage sera biaisé. Par conséquent, le modèle prédictif produit ne sera pas performant, ou du moins, loin d'être optimal.

Pour la visualisation on a utilisé des box plots.

La suppression s'est faite par la méthode de l'intervalle interquartile (IQR)

Chaque ensemble de données peut être divisé en quartiles. Le premier quartile indique que 25 % des points de données sont inférieurs à cette valeur, tandis que le deuxième quartile est considéré comme le point médian de l'ensemble de données. La méthode interquartile trouve les valeurs aberrantes sur les ensembles de données numériques en suivant la procédure ci-dessous

- Trouvez le premier quartile, Q1.
- Trouvez le troisième quartile, Q3.
- Calculez l'IQR. $IQR = Q3 - Q1$.
- Définissez la plage de données normale avec la limite inférieure comme $Q1 - 1.5 * IQR$ et la limite supérieure comme $Q3 + 1.5 * IQR$.
- Tout point de données en dehors de cette plage est considéré comme aberrant et doit être supprimé pour une analyse plus approfondie.

Le concept de quartiles et d'IQR peut être mieux visualisé à partir de la boîte à moustaches. Il a le point minimum et maximum défini comme $Q1 - 1.5 * IQR$ et $Q3 + 1.5 * IQR$ respectivement. Tout point en dehors de cette plage est aberrant.

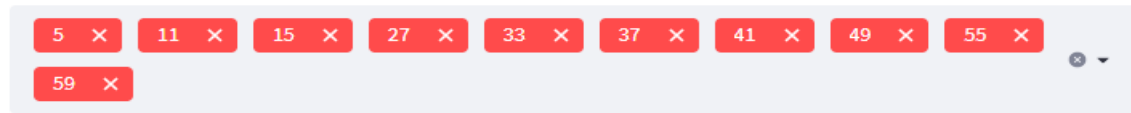
Choisir les données à représenter par des box plots pour visualiser les outliers afin de les supprimer.



Après cette visualisation juste en bas de la page, les lignes qui sont des outliers sont affichés et il y'a un bouton « Remove » cliqué dessus si vous voulez supprimer les outliers.

Outliers

Outliers Index



Remove

4. Features selection

Features selection est le processus de réduction du nombre de variables d'entrée lors du développement d'un modèle prédictif.

Il est souhaitable de réduire le nombre de variables d'entrée à la fois pour réduire le coût de calcul de la modélisation et, dans certains cas, pour améliorer les performances du modèle.

Les méthodes de sélection de variables basées sur les statistiques impliquent l'évaluation de la relation entre chaque variable d'entrée et la variable cible (corrélation) à l'aide de statistiques et la sélection des variables d'entrée qui ont la relation la plus forte avec la variable cible. Ces méthodes peuvent être rapides et efficaces, bien que le choix des mesures statistiques dépende du type de données des variables d'entrée et de sortie.

Pour notre cas on a utilisé principale 2 approches, la première consiste à supprimer les variables qui n'ont pas une forte corrélation avec le Target, la deuxième consiste à évaluer l'apport de chaque feature lors de l'entraînement afin de garder les features qui ont une forte contribution.

Encodage :

Les modèles d'apprentissage automatique exigent que toutes les variables d'entrée et de sortie soient numériques. Cela signifie que si vos données contiennent des données catégorielles, vous devez les coder en nombres avant de pouvoir ajuster et évaluer un modèle.

Les deux techniques les plus populaires sont un codage ordinal et un codage One-Hot.

Pour notre application seule l'encodage ordinal est utilisé.

Dans l'encodage ordinal, chaque valeur de catégorie unique se voit attribuer une valeur entière.

Par exemple, "rouge" est 1, "vert" est 2 et "bleu" est 3.

C'est ce qu'on appelle un codage ordinal ou un codage entier et est facilement réversible. Souvent, des valeurs entières commençant à zéro sont utilisées.

Pour certaines variables, un encodage ordinal peut suffire. Les valeurs entières ont une relation ordonnée naturelle entre elles et les algorithmes d'apprentissage automatique peuvent être en mesure de comprendre et d'exploiter cette relation.

Features Selection

Drop columns

Sélectionner une ou plusieurs colonnes

Choose an option

Encodage

Sélectionner les colonnes à encoder

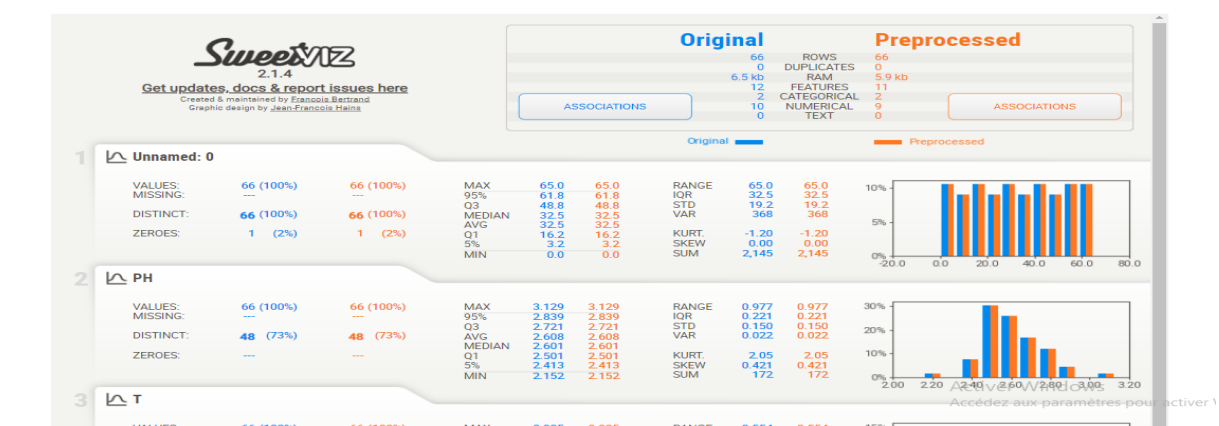
No options to select.

**Linear correlation **

5. Compare

Jusque-là nous avons fait une description du dataset, on l'a traité en supprimant les valeurs manquantes et outliers, on a du features selection et si maintenant on comparait notre dataset origine avec le dataset prétraité.

L'objectif est de faire une comparaison détaillé entre original et celui prétraité, on pourra bien voir toutes les modification qu'on a eu a apporté sur le dataset, si son est satisfait des résultats on pourra continuer sinon on revient faire d'autre traitement



6. La normalisation des données

Normalisation : La normalisation est une méthode de prétraitement des données qui permet de réduire la complexité des modèles.

Elle permet de donner la même importance à tous les features

Il existe plusieurs méthodes à l'instar de Standard scaling.

Standard scaling

La normalisation standardise la moyenne et l'écart-type de tout type de distribution de données, ce qui permet de simplifier le problème d'apprentissage en s'affranchissant de ces deux paramètres.

Pour effectuer cette transformation, on soustrait aux données leur moyenne empirique μ et on les divise par leur écart-type σ .

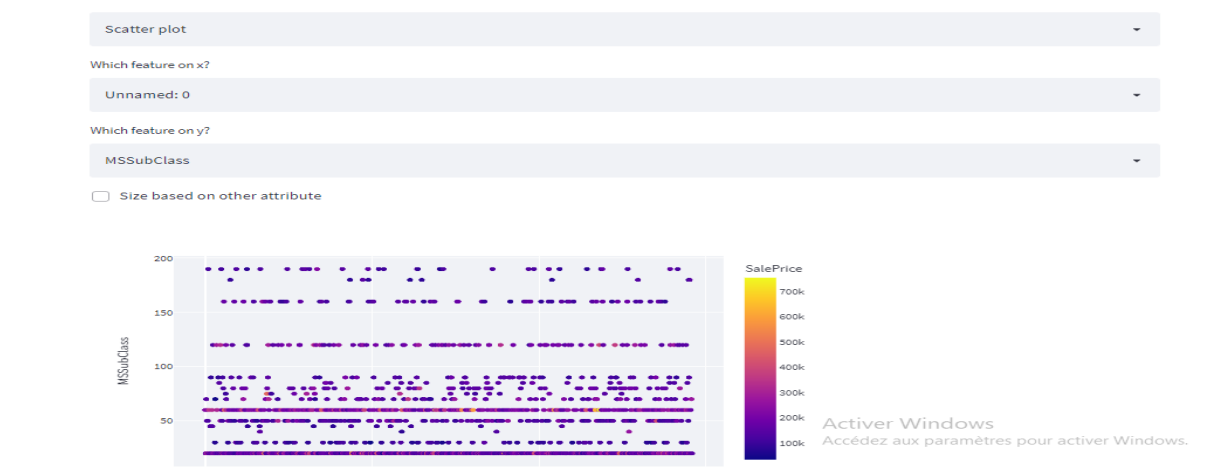
III. Data visualisation

La data visualisation est l'art et la manière de transformer la donnée en un formidable outil d'analyse. En montrant l'invisible, la data visualisation facilite et accélère la prise de décision.

La data visualisation simplifie la diffusion de l'information. Elle apporte des points de comparaison et d'analyse sur les tendances. Elle affine alors les prédictions sur les tendances à venir.

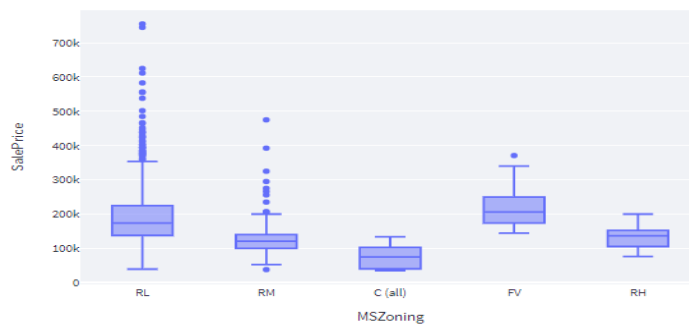
1. Scatter plot

Un nuage de points (également appelé nuage de points ou diagramme de dispersion) est un type de tracé ou de diagramme utilisant des coordonnées généralement deux variables pour un ensemble de données.



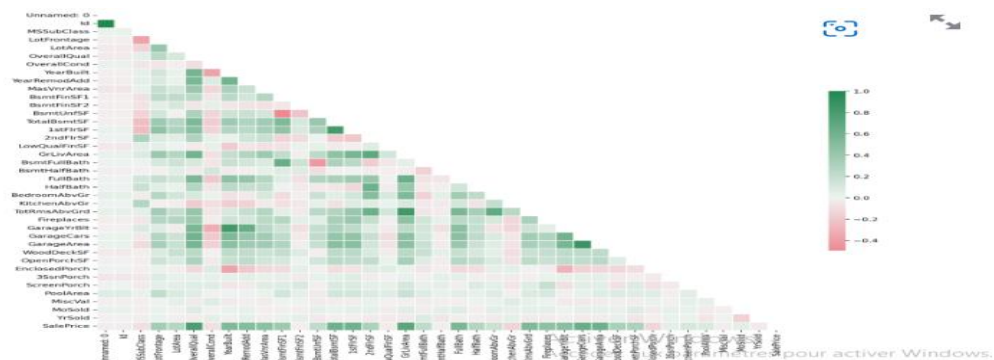
2. Box plot

Un box-plot est un graphique simple composé d'un rectangle duquel deux droites sortent afin de représenter certains éléments des données. La valeur centrale du graphique est la médiane (il existe autant de valeur supérieure qu'inférieures à cette valeur dans l'échantillon).



3. Corrélation matrix

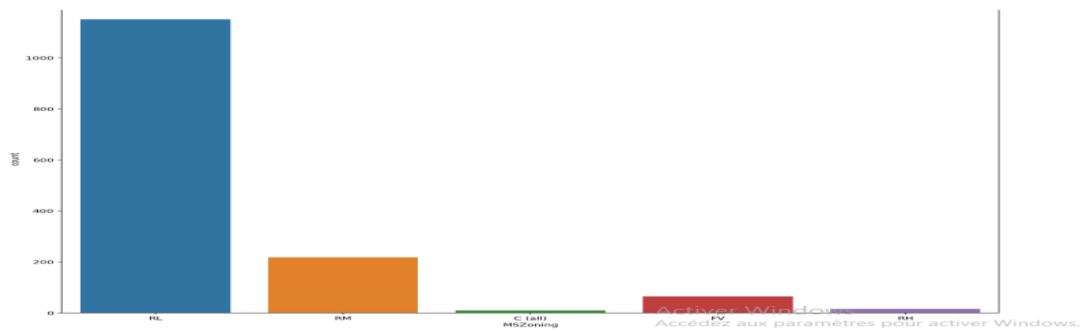
Une matrice de corrélation est simplement un tableau qui affiche les coefficients de corrélation pour différentes variables. La matrice représente la corrélation entre toutes les paires de valeurs possibles dans un tableau. C'est un outil puissant pour résumer un grand ensemble de données et pour identifier et visualiser des modèles dans les données données.



4. Count plot

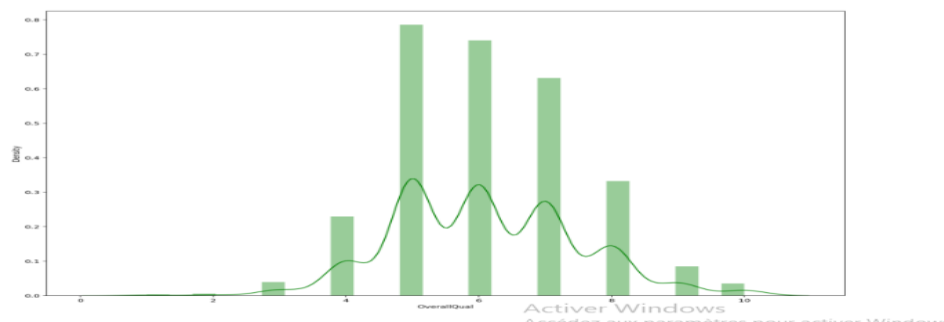
Le count plot est utilisé pour représenter l'occurrence (les comptes) de l'observation présente dans la variable catégorielle.

Il utilise le concept d'un graphique à barres pour la représentation visuelle.



5. Distribution plot

Les diagrammes de distribution sont d'une importance cruciale pour l'analyse exploratoire des données. Ils nous aident à détecter les valeurs aberrantes et l'asymétrie, ou à obtenir un aperçu des mesures de la tendance centrale (moyenne, médiane et mode).



IV. Machine Learning

Après avoir traité nos données à travers le preprocessing (prétraitement des données), les visualisées dans la partie <visualisation>, on va procéder au machine Learning autrement appelée apprentissage automatique.

Le machine Learning est une technique de programmation informatique qui utilise des probabilités statistiques pour donner aux ordinateurs la capacité d'apprendre par eux-mêmes sans programmation explicite.

Cette partie se divisera principalement en deux volets à savoir :

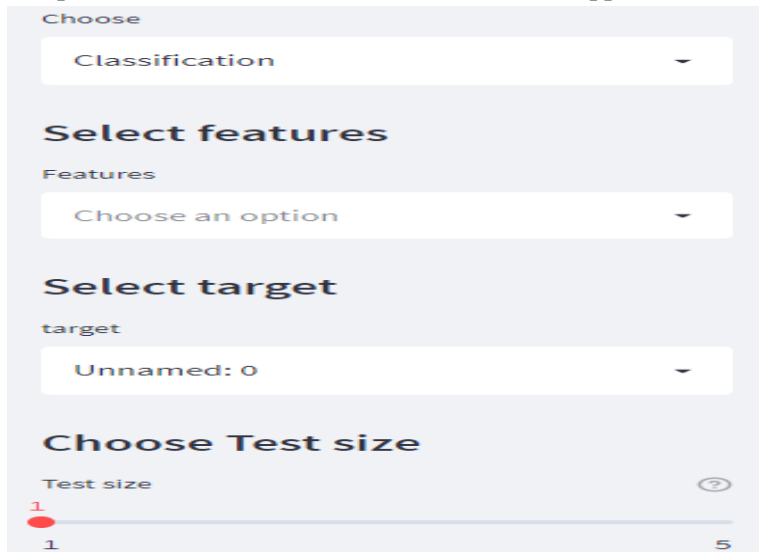
- La classification
- La régression

1. La classification

La classification est une partie du machine Learning où la variable de sortie est catégorie.

Dans cette partie on met en place des modèles, qui une fois bien entraînés, essayeront de développer une fonction qui classifiera avec précision la sortie à partir des variables d'entrées.

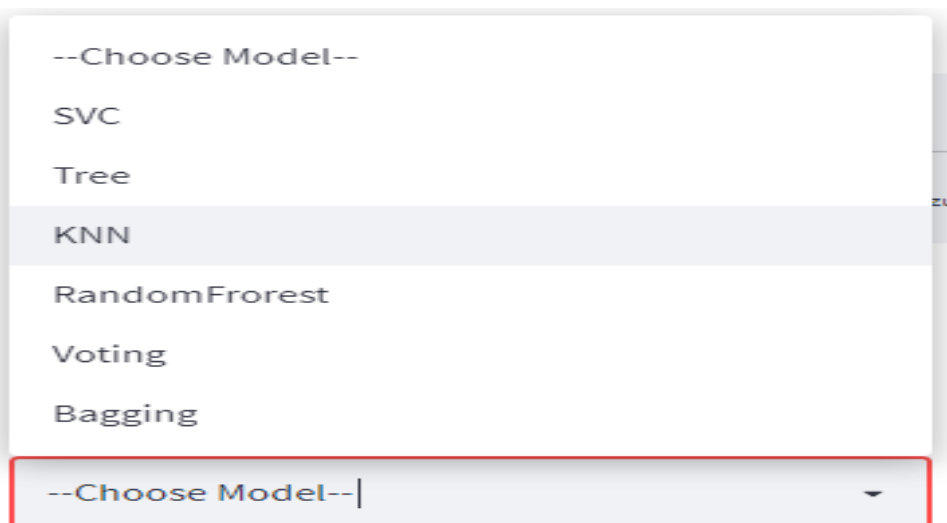
Après avoir choisi la classification il faut choisir les features (variables d'entrées du modèle) et le target, ensuite on choisit le test size, le test size représente le pourcentage du donné de test (test set) et par complémentarité celui du donné d'entraînement appelé le train set.



The screenshot shows a configuration panel for a machine learning model. It includes several sections: 'Choose' with a dropdown set to 'Classification'; 'Select features' with a dropdown set to 'Choose an option'; 'Select target' with a dropdown set to 'Unnamed: 0'; and 'Choose Test size' with a slider ranging from 1 to 5, currently set at 1. A red dot is visible on the slider at the value 1.

Choix du modèle :

Après avoir choisi les features, le target et le test size choisir le modèle qui vous convient entre le SVC (support vector Machine), le Tree, RandomForest, le KNN(k-Neighbors).



The screenshot shows a dropdown menu for selecting a machine learning model. The menu is open, displaying the following options: '--Choose Model--', SVC, Tree, KNN (which is highlighted with a light blue background), RandomForest, Voting, and Bagging. Below the list, there is a red rectangular box highlighting the dropdown control, which contains the text '--Choose Model--' and a downward arrow.

1.a SVC

i. Définition

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support-vector machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de classification et de régression. Les SVM sont une généralisation des classifieurs linéaires.

Les avantages des machines à vecteurs de support sont :

- Efficace dans les espaces de grande dimension.
- Toujours efficace dans les cas où le nombre de dimensions est supérieur au nombre d'échantillons.
- Utilise un sous-ensemble de points d'apprentissage dans la fonction de décision (appelés vecteurs de support), il est donc également efficace en mémoire.
- Polyvalent : différentes fonctions du noyau peuvent être spécifiées pour la fonction de décision. Des noyaux communs sont fournis, mais il est également possible de spécifier des noyaux personnalisés.

Les inconvénients des machines à vecteurs de support incluent :

- Si le nombre de caractéristiques est bien supérieur au nombre d'échantillons, évitez le sur-ajustement dans le choix des fonctions du noyau et le terme de régularisation est crucial.
- Les SVM ne fournissent pas directement des estimations de probabilité, celles-ci sont calculées à l'aide d'une validation croisée quintuple coûteuse.

Après le choix du modèle l'utilisateur devra procéder au paramétrage des hyperparamètres ainsi que l'entraînement.

ii. Paramétrage du modèle (parameters tuning)

On offre la possibilité de choisir les meilleurs hyperparamètres afin que le modèle ait de très bonnes performances.

Choose

Classification

Select features

Features

PH X T X CE X

OZ X NH X NO X

SO X PO X DBO5 X

Select target

target

Classe

Choose Test size

Test size

1 2 5

Model

SVC

Navigation

Auto RandomSearch

SVC Hyper_Params

C:

Min_value C

0,00 - +

Max_value C

0,00 - +

GAMMA:

Min_value Gamma

0,00 - +

Max_value Gamma

0,00 - +

Kernel :

Choose Kernel(s)

Choose an option

Save model

Validator

Active Windows

Accédez aux paramètres pour acti

iii. Résultats

Pour chaque modèle nous allons affiché l'accuracy,F1 score, Recall et la precision.

Accuracy	F1 score	Recall	Precision
0.857	0.714	0.857	0.688
↓ -0.058	↓ -0.216	↓ -0.06	↓ -0.281

Metrics report				
Model Report:				
	precision	recall	f1-score	support
1	0.75	1.00	0.86	3
2	0.00	0.00	0.00	1
3	1.00	1.00	1.00	1
4	1.00	1.00	1.00	2
accuracy			0.86	7
macro avg	0.69	0.75	0.71	7
weighted avg	0.75	0.86	0.80	7

1.b KNN

i. Definition

La classification basée sur les voisins est un type d'apprentissage basé sur *les instances* ou d'apprentissage *non généralisant* : il ne tente pas de construire un modèle interne général, mais stocke simplement des instances des données d'apprentissage. La classification est calculée à partir d'un vote à la majorité simple des plus proches voisins de chaque point : un point de requête se voit attribuer la classe de données qui a le plus de représentants parmi les plus proches voisins du point.

ii. Paramétrage du modèle (parameters tuning)

Choose

Classification

Select features

Features

PH X T X CE X

O2 X NH X NO X

SO X PO X DBOS X

Select target

target

Classe

Choose Test size

Test size

1

1

5

Classification

Model

KNN

Navigation

Auto

GridSearch

KNN Params

n_neighbors

choose min neighbors

1,00

choose max neighbors

5,00

weight

Choose weight

uniform X distance X

Choose algo

ball_tree X auto X kd_tree X

Save model

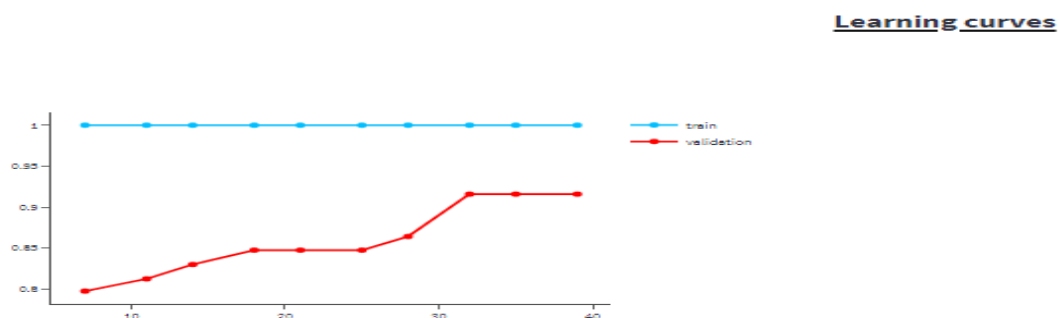
METRICS

iii. Résultat du modèle

Pour ces hyperparametres on obtient les metrics suivants

</

Et des learning curves du train set et du test set



1.c RandomForest Classifier

i. Définition

Les forêts aléatoires ou forêts de décision aléatoires sont une méthode d'apprentissage d'ensemble pour la classification, la régression et d'autres tâches qui fonctionnent en construisant une multitude d'arbres de décision au moment de la formation. Pour les tâches de classification, la sortie de la forêt aléatoire est la classe sélectionnée par la plupart des arbres.

Les forêts de décision aléatoires corrigent l'habitude des arbres de décision de s'adapter à leur ensemble d'entraînement. Les forêts aléatoires surpassent généralement les arbres de décision, mais leur précision est inférieure à celle des arbres à gradient boosté. Cependant, les caractéristiques des données peuvent affecter leurs performances.

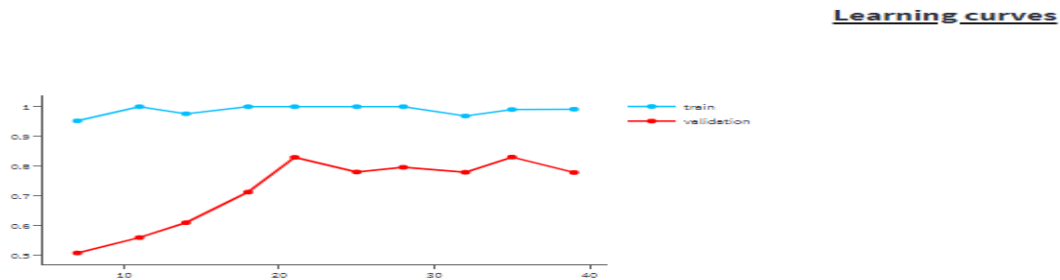
ii. Paramétrage du modèle (parameters tuning)

The interface is titled "Classification" and "RandomForest". It includes a "Choose" dropdown set to "Classification", a "Select features" section with a list of features (PH, T, CE, OZ, NH, NO, SO, PO, DBOS) and a "Select target" section with a dropdown set to "Classe". The "Choose Test size" section shows a slider from 1 to 5. The "RandomForest hyperParams" section includes sliders for "n_estimators" (10 to 20), "min_split" (0,01 to 0,10), "ccp alpha" (0,01 to 0,10), and "max features" (auto, sqrt, log2). A "Save model" button and "Activier Windows" text are also visible.

iii. Résultats obtenus

Les metrics obtenus

Learning curves



1.d Tree

i. Définition

L'arbre de décision est une technique d'apprentissage supervisé qui peut être utilisée à la fois pour les problèmes de classification et de régression, mais elle est généralement préférée pour résoudre les problèmes de classification. Il s'agit d'un classificateur arborescent, où les nœuds internes représentent les caractéristiques d'un ensemble de données, les branches représentent les règles de décision et chaque nœud feuille représente le résultat.

Dans un arbre de décision, il y a deux nœuds, qui sont le nœud de décision et le nœud feuille. Les nœuds de décision sont utilisés pour prendre n'importe quelle décision et ont plusieurs branches, tandis que les nœuds feuilles sont la sortie de ces décisions et ne contiennent pas d'autres branches.

Les décisions ou le test sont effectués sur la base des caractéristiques de l'ensemble de données donné. Il s'agit d'une représentation graphique permettant d'obtenir toutes les solutions possibles à un problème/une décision en fonction de conditions données.

C'est ce qu'on appelle un arbre de décision car, semblable à un arbre, il commence par le nœud racine, qui se développe sur d'autres branches et construit une structure arborescente.

Pour construire un arbre, nous utilisons l'algorithme CART, qui signifie Classification and Regression Tree algorithm.

Un arbre de décision pose simplement une question et, en fonction de la réponse (Oui/Non), il divise ensuite l'arbre en sous-arbres.

ii. Paramétrage du modèle (parameters tuning)

Choose

Classification

Select features

Features

PH

T

CE

O2

NH

NO

SO

PO

DBO5

Select target

target

Classe

Choose Test size

Test size

125

Model

Tree

Navigation

☐ Auto ☒ RandomSearch

Tree Params

min_impurity

choose min value decrease

0,00

-

+

choose max value decrease

0,00

-

+

min_leaf

choose min value

0,00

-

+

choose max value

0,00

-

+

min_split

choose min value split

0,00

-

+

choose max value split

0,00

-

+

ccp_alpha

choose min value ccp

0,00

-

+

choose max value ccp

0,00

-

+

max_features

Choose one

Choose an option

Save model

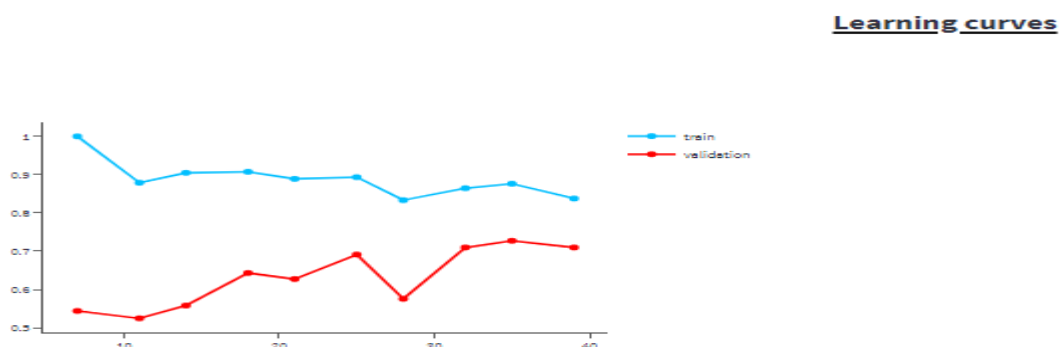
☐ Validar

Activer Windows
Accédez aux paramètres pi

iii. Resultats

METRICS				
Accuracy	F1 score	Recall	Precision	
0.714	0.5	0.714	0.5	
↓ -0.167	↓ -0.202	↑ 0.001	↓ -0.198	
Metrics report				
Model Report:				
	precision	recall	f1-score	support
1	1.00	1.00	1.00	4
2	0.00	0.00	0.00	1
4	0.50	0.50	0.50	2
accuracy			0.71	7
macro avg	0.50	0.50	0.50	7
weighted avg	0.71	0.71	0.71	7

Learning curves



1.e Voting

Définition

Le vote est une méthode permettant à un groupe, tel qu'une assemblée ou un électorat, de prendre une décision collective ou d'exprimer une opinion généralement à la suite de discussions, de débats ou de campagnes électorales.

C'est le même principe qu'on applique pour les modèles.

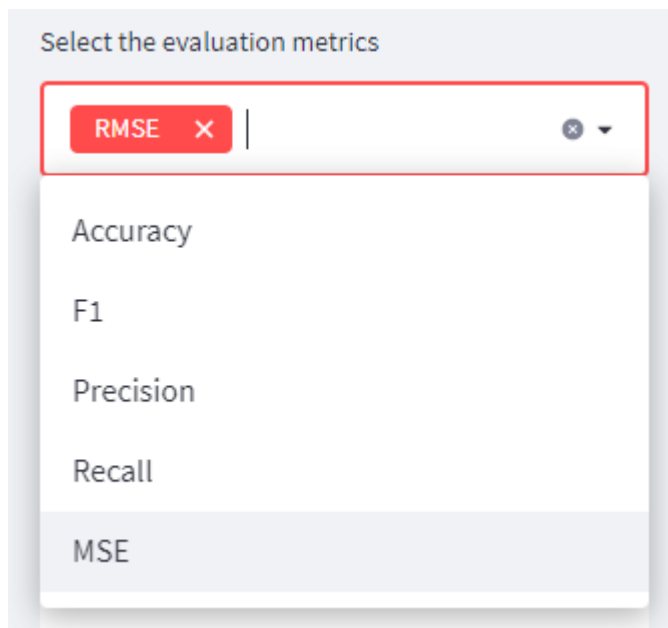
2. Régression

Dans cette partie on offre à l'utilisateur la possibilité de diviser ses données : un pourcentage pour l'entraînement et un autre pour le test.

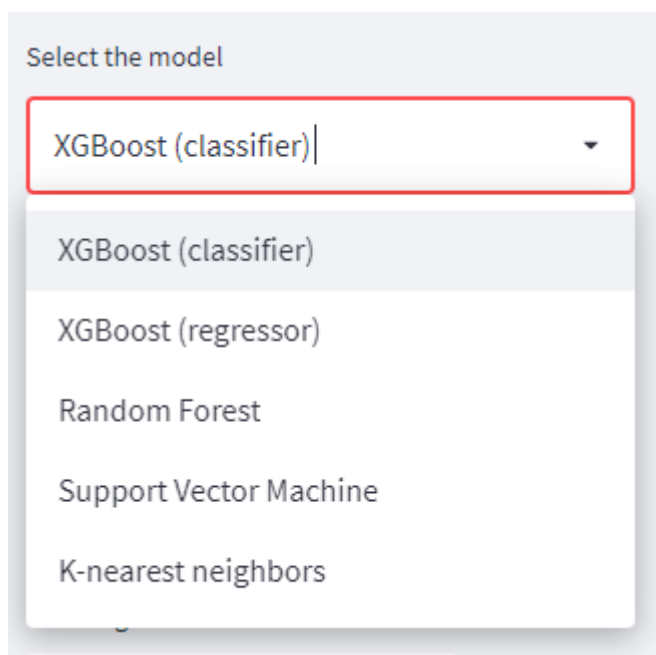


The screenshot shows a web interface with a light gray background. At the top, there is a section titled "Choose" in a small, dark font. Below this title is a white rectangular box with a red border, containing the text "Regression" and a small downward arrow on the right side. Below this box, the text "Experiment parameters:" is displayed in a bold, dark font. Underneath, the label "Test set size" is shown in a small, dark font. Below the label is a horizontal slider bar. The slider bar has a red line and a red circular handle. The value "0.20" is displayed in red text above the handle. The slider bar has two numerical labels at its ends: "0.01" on the left and "0.99" on the right, both in a small, dark font.

A côté de cela on offre également la possibilité de choisir la métrique pour évaluer la performance du modèle de prédiction.



Choisir le modèle



Par ailleurs on peut également paramétrer le modèle comme dans la partie classification.

Model parameters:

Learning rate

0,10 - +

Num. estimators

100

1 500

Maximum depth

3

1 20

Enfin on entraîne le modèle

Running **XGBoost (classifier)** with a test set size of **20%**.

There are **713** instances in the training set and **178** instances in the test set.

Run predictions

☒ Hyperparameters used:

```
{  
  "learning_rate" : 0.1  
  "n_estimators" : 100  
  "max_depth" : 3  
}
```

Warning, only the first metric is selected when using Cross Validation.

Preprocessing completed!

V. Prédiction

Maintenant qu'on a entraîné notre modèle, la dernière étape consistera à faire des tests, prédire la qualité avec notre modèle déjà entraîné.

Pour cela on a 2 options soit prédire la qualité pour une seule mesure ou bien uploadé un fichier de mesure pour prédire la totalité des mesures.

1. Predire pour une seule mesure

Saisir les donner d'entrées

PH

 - +

T

 - +

CE

 - +

O2

 - +

NH

10,00

- +

NO

9,00

- +

SO

8,00

- +

PO

7,00

- +

DBO5

3,00

- +

Et on obtient la prédiction suivante :

La qualité de ton eau est: Excellente

2. Prédire en uploadant un fichier csv

Uploadé le fichier pour la prédiction, choisissez un modèle et enfin lier chaque champs a la colonne correspondante de votre fichier.

Please upload your dataset (CSV format):

Drag and drop file here

Limit 200MB per file • CSV, XLSX

Browse files

Classeur11.csv

4.2 KB

X

Dataset uploaded successfully!

[Home](#)
[Processing](#)
[Visualisation](#)
[MLearning](#)
[Prediction](#)
[Report](#)
[MSP Multiple](#)

Kind of predict

Multiprediction

Model of prediction

knn_0.86

Saisir les donner d'entrées

PH

PH

T

T

CE

CE

O2

O2

NH

NH

NO

NO

Activer Windows

Accédez aux paramètres pour activer Windows.

Après l'opération de prédiction chaque observation est associée à la classe correspondante qu'on insère dans le fichier de prédiction qu'on a uploadé.

Vous pourrez télécharger le fichier

	Unn:	PH	T	CE	O2	NH	NO	SO	PO	DBO5	IQE	Clas:	classes
0	0	2.4080	0.3510	0.0020	4.9400	2.5060	0.0310	0.0240	0.2850	1.9930	17.0060	1	Excellente
1	1	2.4890	0.3890	0.0020	4.3880	2.5060	0.0220	0.0320	0.3130	2.0500	16.1410	1	Excellente
2	2	2.4640	0.4520	0.0020	4.1900	2.3920	0.0360	0.0290	0.3700	2.1530	15.8530	1	Excellente
3	3	2.7110	0.5160	0.0020	4.2130	7.4040	0.0280	0.0310	0.4270	2.3350	21.3640	1	Excellente
4	4	2.5100	0.7060	0.0030	3.6900	5.6950	0.0470	0.0730	42.9990	6.8910	65.0980	4	Mauvaise
5	5	2.3700	0.6740	0.0080	2.0920	301.8510	1.1120	0.0910	55.5580	29.1030	388.5430	3	Peu Polluée
6	6	2.8510	0.3510	0.0020	5.6950	1.3670	0.0130	0.0210	0.2560	0.7970	16.5510	1	Excellente
7	7	2.6260	0.4270	0.0020	4.9690	1.2530	0.0150	0.0230	0.4270	1.3100	15.5690	1	Excellente
8	8	2.6190	0.3960	0.0020	4.8820	0.9110	0.0250	0.0250	0.3130	0.5700	14.1980	1	Excellente
9	9	2.5240	0.6010	0.0020	4.1550	1.3670	0.0280	0.0240	0.4560	2.2900	15.1890	1	Excellente

VI. Le rapport

Cette dernière partie comportera un aperçue claire des modèles entrainés avec leur précision, les meilleurs hyperparametres, l'accuracy ainsi que des graphiques pour la visualisation de ces données.

Welcome

User's Data

	nomModele	hyperpara	PRECISION	RECALL_SCORE
0	svc2022-06-23##19:17:47	{'C': 0.01, 'gamma': 0.01, 'kernel': 'sigmoid'}	0.10714285714285714	0.42857142857142855
1	svc2022-06-23##19:27:58	{'C': 0.01, 'gamma': 0.01, 'kernel': 'poly'}	0.9642857142857143	0.9285714285714286

[Download Report](#)

On a la possibilité de télécharger ces données pour les réutiliser après.

Welcome

User's Data

		PRECISION	RECALL_SCORE	F1_SCORE	ACCURACY
0	kernel: 'sigmoid'}	0.10714285714285714	0.42857142857142855	0.2571428571428571	0.42857142857142855
1	kernel: 'poly'}	0.9642857142857143	0.9285714285714286	0.9273504273504274	0.9285714285714286

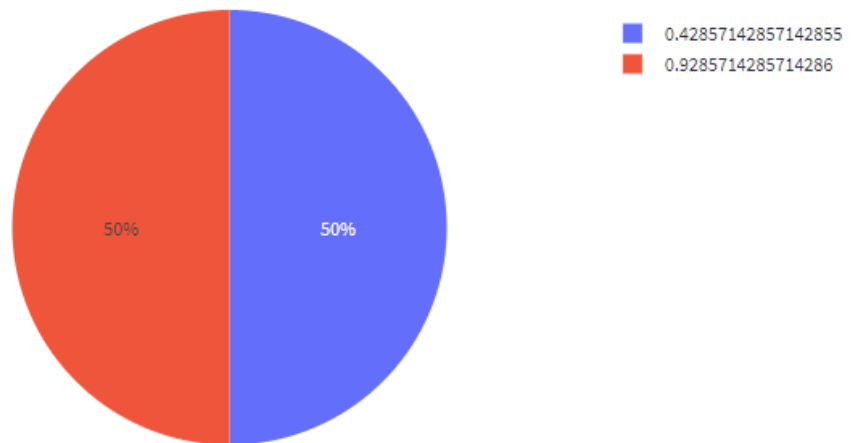
[Download Report](#)





ACCURACY

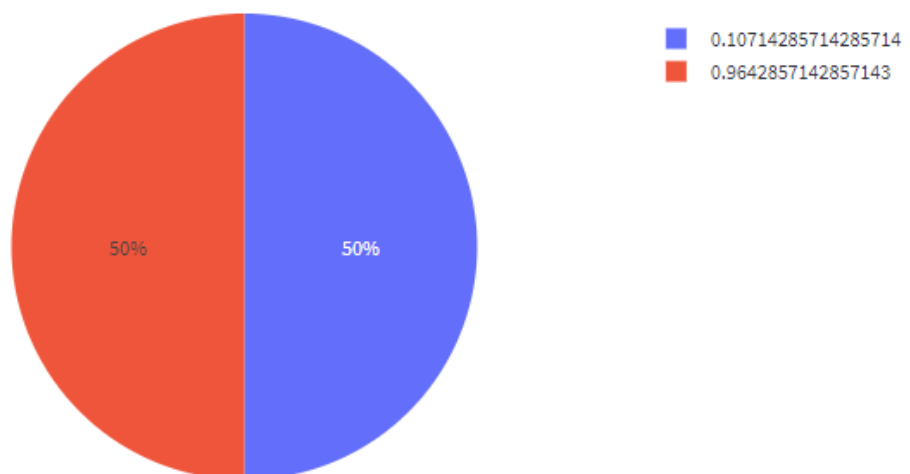
ACCCURITY

Accuracy





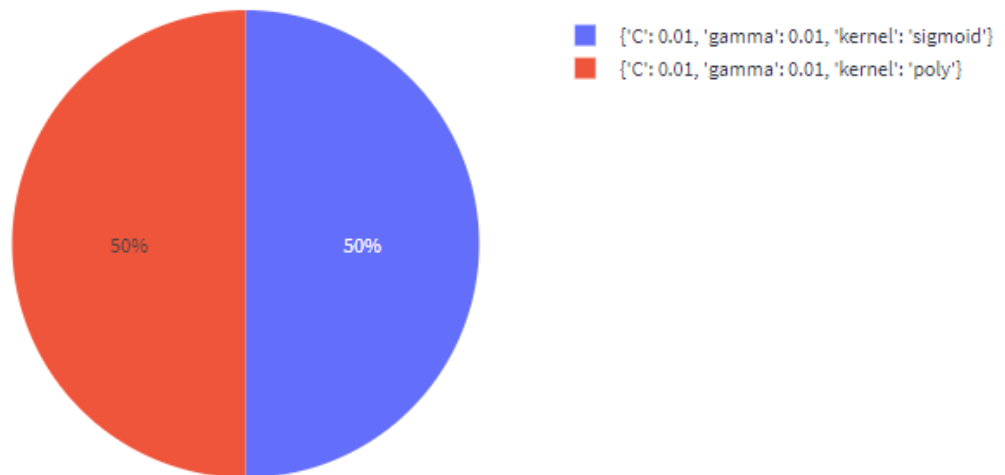
** Pie-Chart for modele's precision**

Pie-Chart  for modele's  precision



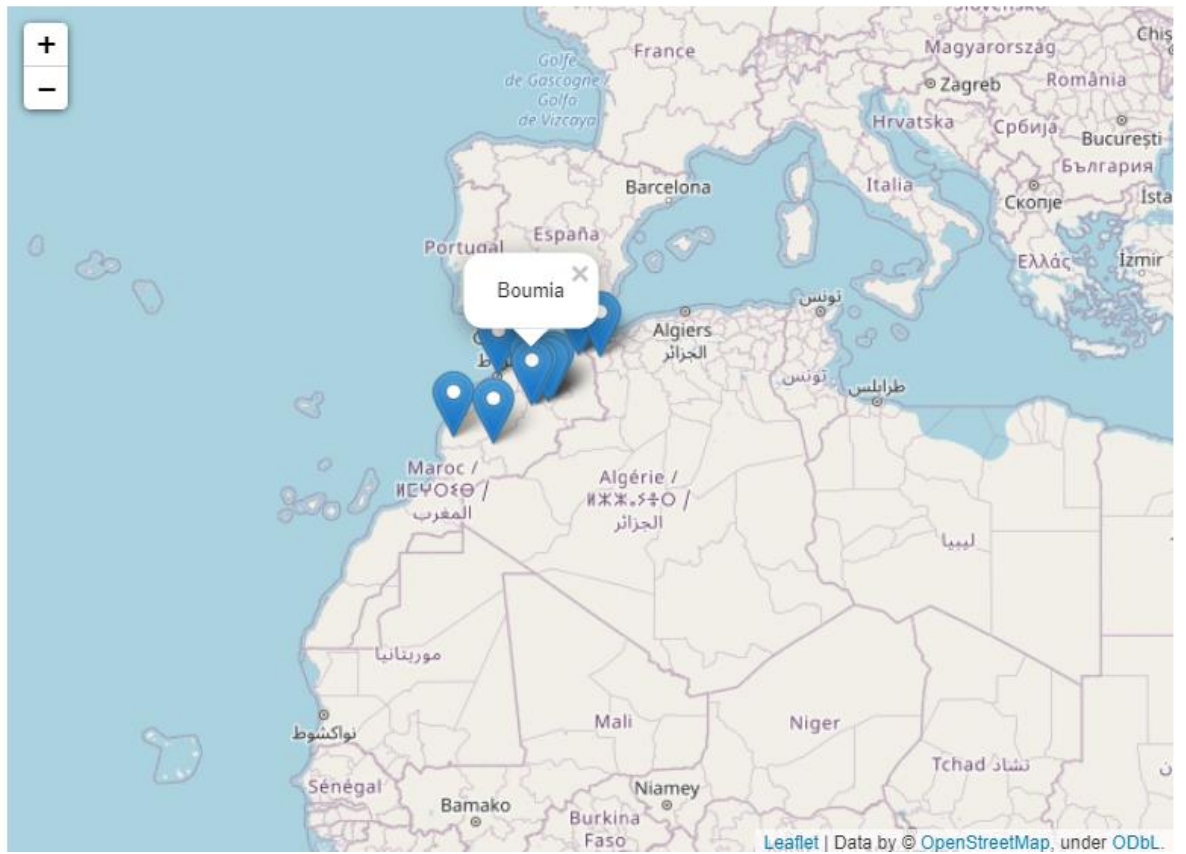
** Pie-Chart for modele's hyperparameters**

Pie-Chart  for modele's  hyperparametres



VII. Le map

Ci-dessous une vue des differentes stations du fleuve MOULOUYA



Conclusion

Ce projet nous ont permis d'avoir une maîtrise en termes de prétraitement des données et visualisation.

Pouvoir télécharger des données, les visualiser par des graphiques concis, détecter et supprimer les valeurs manquantes, le pourcentage de valeurs manquantes, entraîner des modèles d'apprentissage sur des données bien traiter et enfin sauvegarder sous forme de base de données, telles sont les principaux objectifs réalisés de ce projet.

Références

https://github.com/antonin-lfv/Online_preprocessing_for_ML/blob/master/main.py

https://share.streamlit.io/soft-nougat/dqw-ivves_structured/main/app.py

<https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/>

Documentation Streamlit

cours machine learning de Mr BERRADA ensam meknes