



جامعة هواري بومدين للعلوم والتكنولوجيا
Université des Sciences et de la Technologie
Houari Boumediene



Rapport du projet

Module : Data Mining

Intitulé :

Exploitation des données et Extraction des règles d'associations .

Spécialité : Systèmes Informatiques Intelligents

Réalisé par :

- DJENANE Nihad
- M'BAREK Lydia

Travail demandé par :

MME. DRIAS HABIBA - Cours
MME. BELKADI WIDAD HASSINA - TP

Promotion : 2023/2024

Table des matières

Table des figures	ii
Liste des tableaux	iv
Introduction générale	1
1 Analyse et prétraitement des données	2
1.1 Introduction	2
1.2 Données statiques	2
1.2.1 Description Générale du Dataset 1	2
1.2.2 Analyse des données	4
1.2.3 Prétraitement	8
1.3 Données temporelles	15
1.3.1 Description Générale du Dataset 2	15
1.3.2 Prétraitement	16
1.3.3 Visualisation	16
2 Extraction de motifs fréquents, règles d'associations et corrélations	22
2.1 Introduction	22
2.2 Discrétisation des Données : Approches Equal-Frequency et Equal-Width	22
2.2.1 Méthode 1 : Discrétisation en classes d'effectifs égaux	22
2.2.2 Méthode 2 : Discrétisation en classes d'amplitudes égales	22
2.3 Extraction des motifs fréquents puis les règles d'association en utilisant l'algorithme Apriori	23
2.3.1 Extraction des motifs fréquents	23
2.3.2 Extraction des règles d'associations et corrélations	25
2.3.3 Expérimentation	27
Conclusion générale	29

Table des figures

1.1	Quelques instances du Dataset 1	2
1.2	Visualisation du Dataset 1	3
1.3	Description du Dataset 1	3
1.4	BoxPlot attribut N	4
1.5	BoxPlot attribut K	4
1.6	BoxPlot attribut S	5
1.7	BoxPlot attribut Mn	5
1.8	BoxPlot attribut B	5
1.9	Histogramme attribut N	6
1.10	Histogramme attribut K	6
1.11	Histogramme attribut S	7
1.12	Histogramme attribut Mn	7
1.13	Histogramme attribut B	8
1.14	Exemple valeur manquante	8
1.15	BoxPlot2 attribut N	9
1.16	BoxPlot2 attribut K	9
1.17	BoxPlot2 attribut S	9
1.18	BoxPlot2 attribut Mn	9
1.19	BoxPlot2 attribut B	9
1.20	Matrice des coefficients de corrélation	10
1.21	Dispersion 'Fertility' et 'N'	10
1.22	Dispersion 'OC' et 'OM'	11
1.23	Dispersion 'Ph' et 'Zn'	11
1.24	Dispersion 'K' et 'EC'	12
1.25	Dispersion 'Mn' et 'N'	12
1.26	Dispersion 'EC' et 'Zn'	13
1.27	Dataset après réduction H	13
1.28	Dataset après réduction V	14
1.29	Normalisation min-max	14
1.30	Normalisation z-score	15
1.31	Quelques instances du Dataset 2	15
1.32	Dataset 2	16
1.33	Distribution	17
1.34	Evolution hebdomadaire	17
1.35	Evolution mensuel	18
1.36	Evolution annuel	18
1.37	Distribution cas covid positive	19
1.38	Rapport population et test	19

1.39	Distribution testes positives	20
1.40	Rapport p=46	20
1.41	Rapport p=47	20
1.42	Rapport p=33	21
1.43	Rapport p=62	21
1.44	Rapport p=56	21
1.45	Rapport p=37	21
2.1	Quelques instances du Dataset 3	24
2.2	Motifs Fréquents Extraits avec un Support Minimum de 0.5	24
2.3	Règles d'Association Générées avec une Confiance Minimale de 0.5	26

Liste des tableaux

1.1	Tendances centrales des caractéristiques du dataset 1	4
-----	---	---

Introduction générale

Le Data Mining est actuellement appliqué dans divers domaines tels que la recherche, le marketing, l'éducation, la finance et la santé. Sa popularité réside dans sa capacité à analyser d'importantes quantités d'informations et à résoudre rapidement des problèmes. La qualité des résultats obtenus est étroitement liée à la qualité des données manipulées. Étant donné que les données du monde réel sont souvent brutes, il est impératif d'effectuer un prétraitement des données pour garantir une pertinence et une fiabilité optimales.

Dans ce contexte, le premier chapitre de notre rapport se consacre à l'analyse de deux types de données distincts. D'une part, des données statiques provenant d'un ensemble de données englobant des informations détaillées liées à la qualité du sol. D'autre part, des données temporelles documentant l'impact de la pandémie de COVID-19 dans plusieurs zones. L'objectif central est d'appliquer des techniques avancées de Data Mining pour visualiser et identifier des tendances ainsi que des corrélations significatives.

Dans le deuxième chapitre, notre objectif a été d'extraire des informations et des règles d'association en exploitant une dataset contenant des données cruciales liées aux conditions environnementales dans le contexte agricole. Nous avons utilisé l'algorithme Apriori pour favoriser une compréhension approfondie des relations entre les divers éléments.

Chapitre 1

Analyse et prétraitement des données

1.1 Introduction

L'analyse des données fait référence à l'analyse des types d'attributs qui composent nos données, des types de valeurs de chaque attribut, de la nature des attributs (discrets ou continus), de la distribution des valeurs, etc. Une fois l'analyse des données complétée, nous procédons au prétraitement des données afin d'éliminer le bruit, de corriger les incohérences et les valeurs aberrantes. L'objectif de ce prétraitement est de garantir la qualité des données, de les adapter aux exigences des algorithmes, facilitant ainsi une exploration efficace en vue d'extraire des informations pertinentes et fiables.

Dans ce chapitre, nous allons analyser deux types de données : des données statiques et des données temporelles.

1.2 Données statiques

1.2.1 Description Générale du Dataset 1

Le Dataset 1 englobe des informations détaillées liées à la qualité du sol, comprenant 885 instances. Ces données, exclusivement numériques, encapsulent des mesures telles que les concentrations de nutriments essentiels (N, P, K), le pH, la conductivité électrique (EC), et des éléments tels que le zinc (Zn), le fer (Fe), le cuivre (Cu), le manganèse (Mn), le bore (B), et la matière organique (OM). La colonne 'Fertility' quant à elle, offre une classification de la fertilité du sol.

Ci-dessous, vous trouverez une capture d'écran illustrant quelques instances sélectionnées de notre dataset.

1	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
2	138	8.6	560	7.46	0.62	0.7	5.9	0.24	0.31	0.77	8.71	0.11	1.204	0
3	213	7.5	338	7.62	0.75	1.06	25.4	0.3	0.86	1.54	2.89	2.29	1.8232	0
4	163	9.6	718	7.59	0.51	1.11	14.3	0.3	0.86	1.57	2.7	2.03	1.9092	0
5	157	6.8	475	7.64	0.58	0.94	26	0.34	0.54	1.53	2.65	1.82	1.6168	0
6	270	9.9	444	7.63	0.4	0.86	11.8	0.25	0.76	1.69	2.43	2.26	1.4792	1
7	220	8.6	444	7.43	0.65	0.72	11.7	0.37	0.66	0.9	2.19	1.82	1.2384	0
8	220	7.2	222	7.62	0.43	0.81	7.4	0.34	0.69	1.05	2	1.88	1.3932	0
9	207	7	401	7.63	0.59	0.69	7.6	0.32	0.68	0.62	2.43	1.68	1.1868	0
10	333	14.9	422	8.26	0.48	NA	8.45	0.51	3.32	1.08	9.21	0.32	2.0124	2

FIGURE 1.1 – Quelques instances du Dataset 1

Pour pouvoir visualiser le Dataset 1, on a utilisé la bibliothèque PANDAS qui donne une vue de data comme suit :

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
0	138	8.6	560	7.46	0.62	0.70	5.90	0.24	0.31	0.77	8.71	0.11	1.2040	0
1	213	7.5	338	7.62	0.75	1.06	25.40	0.30	0.86	1.54	2.89	2.29	1.8232	0
2	163	9.6	718	7.59	0.51	1.11	14.30	0.30	0.86	1.57	2.70	2.03	1.9092	0
3	157	6.8	475	7.64	0.58	0.94	26.00	0.34	0.54	1.53	2.65	1.82	1.6168	0
4	270	9.9	444	7.63	0.40	0.86	11.80	0.25	0.76	1.69	2.43	2.26	1.4792	1
5	220	8.6	444	7.43	0.65	0.72	11.70	0.37	0.66	0.90	2.19	1.82	1.2384	0
6	220	7.2	222	7.62	0.43	0.81	7.40	0.34	0.69	1.05	2.00	1.88	1.3932	0
7	207	7	401	7.63	0.59	0.69	7.60	0.32	0.68	0.62	2.43	1.68	1.1868	0
8	333	14.9	422	8.26	0.48	NaN	8.45	0.51	3.32	1.08	9.21	0.32	2.0124	2
9	289	8.6	560	7.58	0.44	0.67	7.30	0.63	0.66	0.94	2.43	1.79	1.1524	1

FIGURE 1.2 – Visualisation du Dataset 1

Le Dataset 1 se compose de 885 lignes et 14 colonnes, chaque colonne correspondant à un attribut spécifique tel que N, PK, PH, EC, OC, S, Zn, Fe, Cu, Mn, B, OM, Fertility. Chaque attribut est associé à son propre type de données, qu'il s'agisse de nombres entiers, de nombres décimaux ou de chaînes de caractères. Une capture d'écran illustrant les résultats de la fonction de description du Dataset 1 est présentée ci-dessous.

```
la taille de data est : (885, 14)

le type de 1 attribut est: <class 'numpy.int64'>

le type de 2 attribut est: <class 'str'>

le type de 3 attribut est: <class 'numpy.int64'>

le type de 4 attribut est: <class 'numpy.float64'>

le type de 5 attribut est: <class 'numpy.float64'>

le type de 6 attribut est: <class 'numpy.float64'>

le type de 7 attribut est: <class 'numpy.float64'>

le type de 8 attribut est: <class 'numpy.float64'>

le type de 9 attribut est: <class 'numpy.float64'>

le type de 10 attribut est: <class 'numpy.float64'>
```

FIGURE 1.3 – Description du Dataset 1

1.2.2 Analyse des données

Dans cette étape d'analyse des caractéristiques des attributs, nous avons procédé au calcul des tendances centrales pour chaque attribut, permettant ainsi de déterminer les mesures de dispersion et de symétrie, et ces résultats sont avant le prétraitement de dataset 1.

Dans le tableau suivant on a les tendances centrales pour chaque attribut du Dataset 1 :

	N	P	K	pH	EC	OS	S	Zn	Fe	Cu	Mn	B	OM	Fertility
Moyenne	247	14.55	501.34	7.51	0.54	0.62	7.55	0.47	4.13	0.95	8.65	0.59	1.06	0.59
Médiane	257	8.1	475	7.5	0.55	0.59	6.64	0.36	3.56	0.93	8.34	0.41	1.01	1
Mode	207	8.3	444	7.5	0.62	0.88	4.22	0.28	6.32	1.25	7.54	0.34	1.51	1

TABLE 1.1 – Tendances centrales des caractéristiques du dataset 1

Après avoir examiné le tableau des tendances centrales, nous pouvons identifier les attributs présentant des symétries, à savoir : PH

À présent, nous créons des boxplots pour chaque attribut, incluant les données aberrantes. Nous présentons quelques exemples de boxplots pour les attributs suivants :

— **Attribut N :**

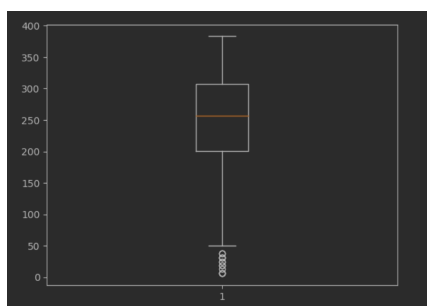


FIGURE 1.4 – BoxPlot attribut N

Observation : Dans cette box plot, on observe la distribution des données avec une ligne médiane, les quartiles (les bords supérieur et inférieur de la boîte), et des points au-delà des "moustaches" qui représentent des valeurs aberrantes. Les valeurs aberrantes, situées au début de la distribution (avant Q1), sont des observations qui s'écartent significativement de la tendance générale des données. Ces points inhabituels sont visualisés en dehors des limites définies par les moustaches.

— **Attribut K :**

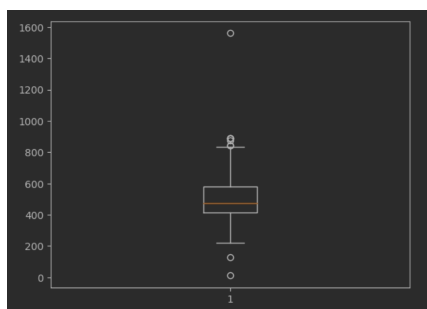


FIGURE 1.5 – BoxPlot attribut K

Observation : Dans cette box plot, les valeurs aberrantes sont situées à la fois au début de la distribution (avant Q1) et à la fin (après Q3).

— **Attribut S :**

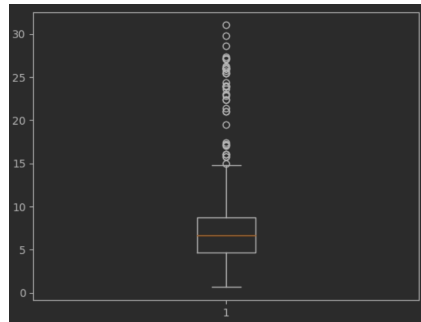


FIGURE 1.6 – BoxPlot attribut S

Observation : Notamment, dans cette box plot, on remarque un nombre significatif de valeurs aberrantes à la fin de la distribution (après Q3).

— **Attribut Mn :**

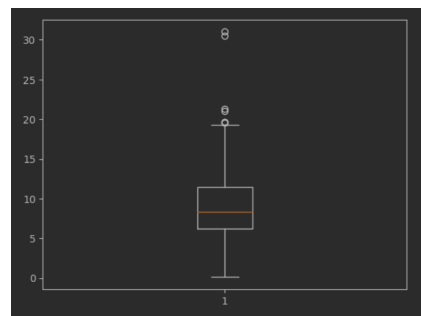


FIGURE 1.7 – BoxPlot attribut Mn

Observation : Dans cette box plot, on remarque quelques valeurs aberrantes à la fin de la distribution (après Q3).

— **Attribut B :**

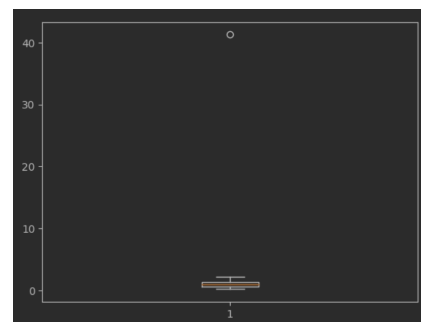


FIGURE 1.8 – BoxPlot attribut B

Observation : Dans cette box plot, la présence d'outliers est quasiment nulle, indiquant une distribution des données relativement homogène et équilibrée.

Suite à l'analyse des boxplots, nous générons les histogrammes des attributs afin de visualiser la distribution des données. Nous présentons quelques exemples d'histogrammes pour les attributs suivants :

— **Attribut N :**

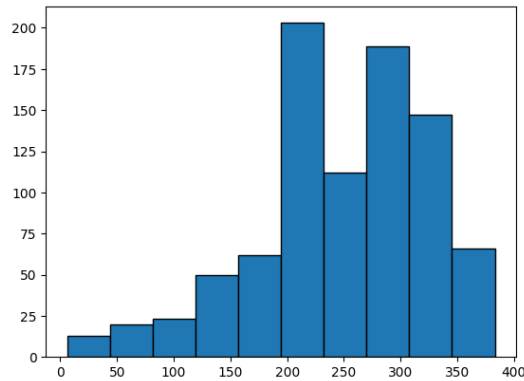


FIGURE 1.9 – Histogramme attribut N

Observation : L'histogramme met en évidence la répartition des valeurs de l'attribut "N" dans le dataset. On observe une concentration significative autour de 200.

— **Attribut K :**

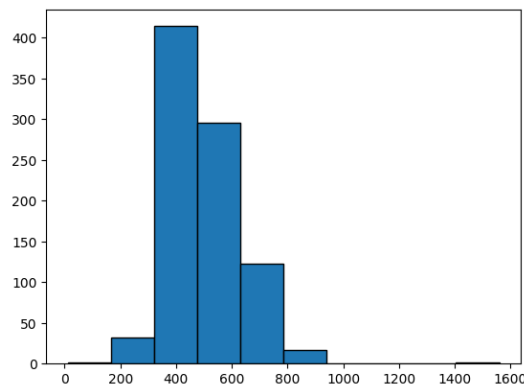


FIGURE 1.10 – Histogramme attribut K

Observation : L'histogramme met en évidence la répartition des valeurs de l'attribut "K" dans le dataset. On observe une concentration significative autour de 400, après la valeur 800, on remarque une absence de barre dans l'histogramme, indiquant une rareté des occurrences au-delà de cette limite.

— **Attribut S :**

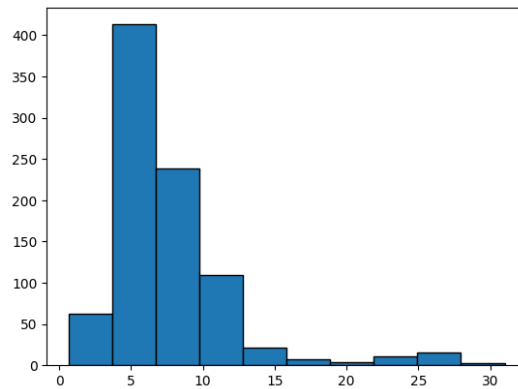


FIGURE 1.11 – Histogramme attribut S

Observation : L'histogramme met en évidence la répartition des valeurs de l'attribut "S" dans le dataset. On observe une concentration significative autour de 5, après la valeur 15, on remarque une absence de barre dans l'histogramme, indiquant une rareté des occurrences au-delà de cette limite

— **Attribut Mn :**

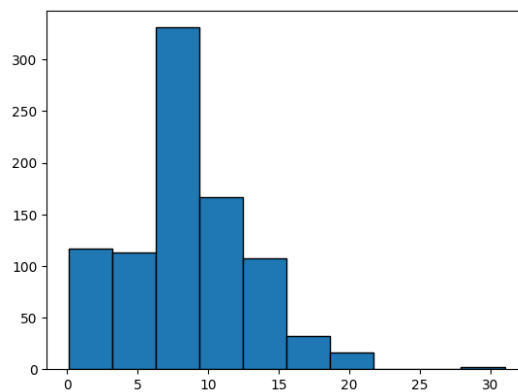


FIGURE 1.12 – Histogramme attribut Mn

Observation : L'histogramme met en évidence la répartition des valeurs de l'attribut "Mn" dans le dataset. On observe une concentration significative autour de 5 et 10, après la valeur 20, on remarque une absence de barre dans l'histogramme, indiquant une rareté des occurrences au-delà de cette limite.

— **Attribut B :**

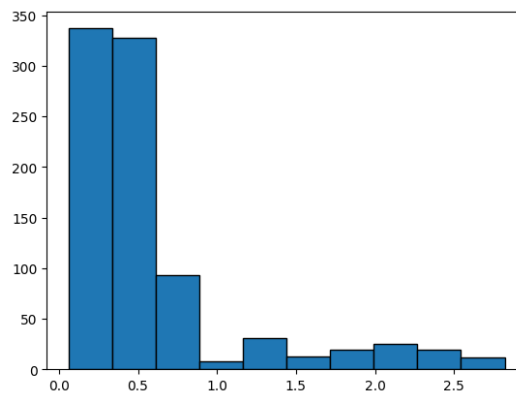


FIGURE 1.13 – Histogramme attribut B

Observation : L'histogramme met en évidence la répartition des valeurs de l'attribut "B" dans le dataset. On observe une concentration significative autour de 0 jusqu'à 0.5, après la valeur 1, on remarque que les occurrences diminuent.

1.2.3 Prétraitement

Après l'analyse des données, nous procédons au prétraitement dans le but de préparer ces données brutes afin de les utiliser pour l'extraction d'informations ou l'entraînement des modèles d'apprentissage automatique. Dans cette section, nous allons nous intéresser au processus de prétraitement suivant :

1. Traitement des valeurs manquantes et aberrantes

Les valeurs manquantes et aberrantes influent sur la qualité des données, pouvant induire des résultats erronés. Afin de remédier à cela, nous allons explorer différentes méthodes de traiter ces valeurs :

- Suppression des valeurs manquantes et aberrantes
- Remplacement des valeurs manquantes par les mesures de tendance centrale.

Dans notre contexte, nous avons débuté par le calcul du nombre de valeurs manquantes pour chaque attribut, comme illustré dans l'exemple ci-dessous :

```
Pour l'attribut 2 P:  
Nombre de valeurs manquantes : 2  
Pourcentage de valeurs manquantes : 0.22598870056497175%
```

FIGURE 1.14 – Exemple valeur manquante

Après ce calcul, nous avons choisis de remplacer les valeurs manquantes par le mode de leur attribut.

En ce qui concerne les valeurs aberrantes, bien que nous ayons créé une fonction pour les supprimer de chaque attribut, nous avons constaté que dans certains cas, des valeurs aberrantes subsistent. Pour remédier à cela, nous avons développé une fonction de détection des outliers, suivie d'une boucle while qui appelle la fonction de suppression tant

que des valeurs aberrantes sont détectées, assurant ainsi un nettoyage complet des données.

Visualisation des Boxplot après le prétraitement

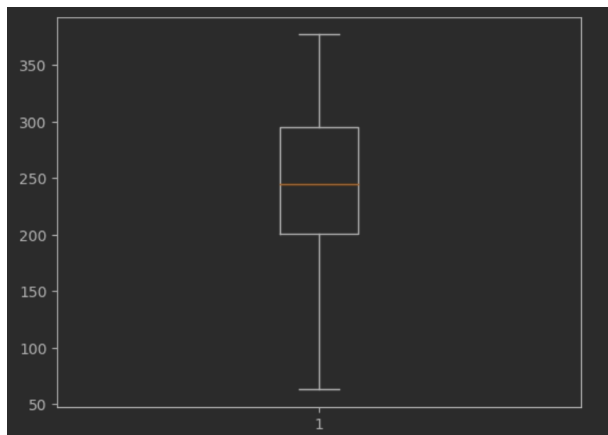


FIGURE 1.15 – BoxPlot2 attribut N

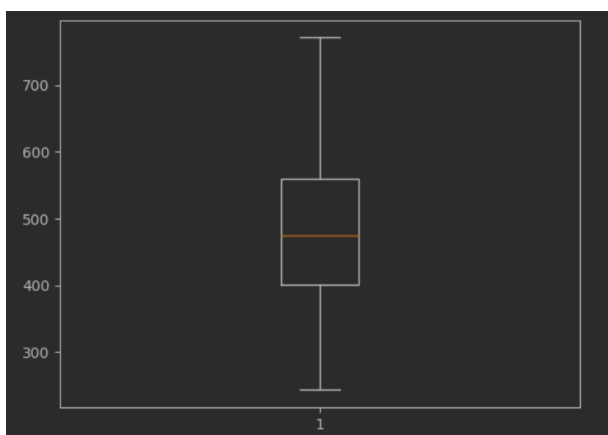


FIGURE 1.16 – BoxPlot2 attribut K

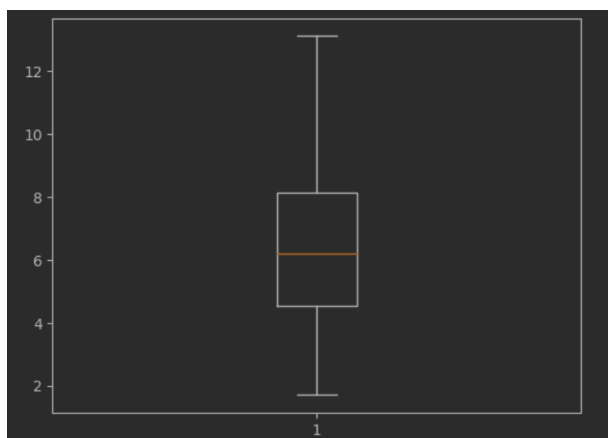


FIGURE 1.17 – BoxPlot2 attribut S

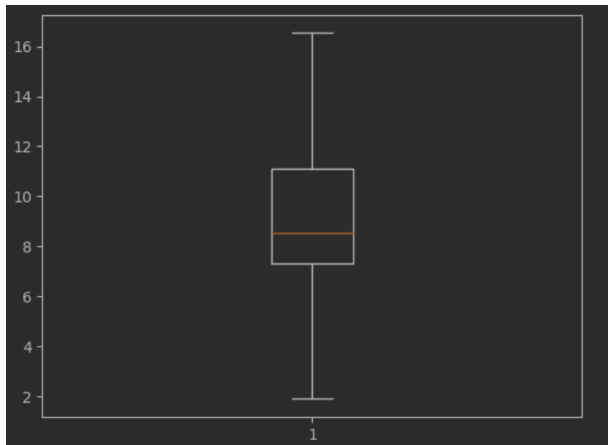


FIGURE 1.18 – BoxPlot2 attribut Mn

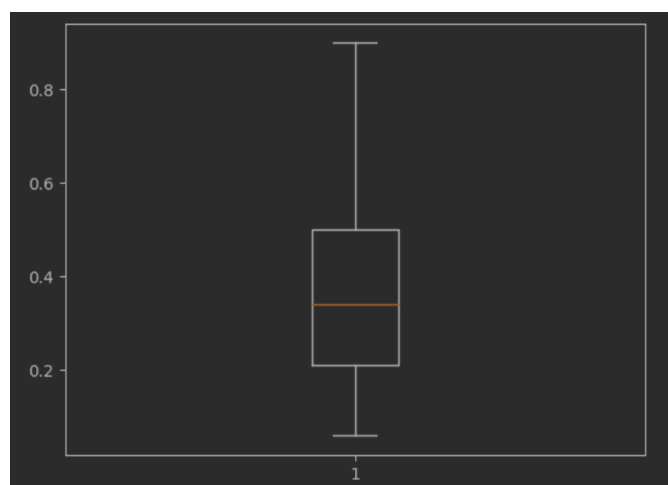


FIGURE 1.19 – BoxPlot2 attribut B

Après la révision des boxplots suite au prétraitement du dataset1, nous abordons maintenant le calcul des coefficients de corrélation de Pearson entre les attributs.

À cet effet, nous avons créé une fonction *CORRELATION* qui calcule le coefficient de corrélation de Pearson entre deux attributs. Ensuite, une fonction *MATRICE CORRELATION* est mise en place pour appeler la fonction *CORRELATION* pour toutes les combinaisons d'attributs, aboutissant ainsi à une matrice de dimensions 14*14 présentant les résultats.

	N	P	K	pH	EC	OC	S	Zn	Fe
N	NaN	0.218244	0.072211	0.108653	0.012409	0.013729	-0.006725	-0.014125	0.012483
P	0.218244	NaN	0.100949	0.006837	0.045107	0.015636	-0.020273	0.05504	0.068479
K	0.072211	0.100949	NaN	-0.099046	-0.139368	-0.02812	0.039522	0.146724	0.02208
pH	0.108653	0.006837	-0.099046	NaN	0.236424	-0.029173	-0.027182	-0.099292	-0.034018
EC	0.012409	0.045107	-0.139368	0.236424	NaN	-0.024898	0.01141	-0.000915	-0.030419
OC	0.013729	0.015636	-0.02812	-0.029173	-0.024898	NaN	0.061294	0.094273	0.07336
S	-0.006725	-0.020273	0.039522	-0.027182	0.01141	0.061294	NaN	0.154296	0.141451
Zn	-0.014125	0.05504	0.146724	-0.099292	-0.000915	0.094273	0.154296	NaN	0.273374
Fe	0.012483	0.068479	0.02208	-0.034018	-0.030419	0.07336	0.141451	0.273374	NaN
Cu	-0.030157	0.016892	0.054588	0.008509	0.056659	0.077807	0.14565	0.18759	0.243525
Mn	-0.000499	-0.021004	0.058203	-0.089554	-0.057156	-0.004585	0.12753	0.129721	0.241028
B	0.105924	0.111242	0.25681	-0.098603	-0.030374	0.104963	0.133238	0.300696	0.161845
OM	0.013729	0.015636	-0.02812	-0.029173	-0.024898	1.002037	0.061294	0.094273	0.07336
Fer...	0.785581	0.278513	0.021366	0.122183	0.016007	0.025536	0.015847	-0.024946	0.010807

FIGURE 1.20 – Matrice des coefficients de corrélation

Après l'analyse des coefficients de corrélation, nous procédons désormais à la création des diagrammes de dispersion des données. Vous trouverez ci-dessous quelques exemples illustratifs.

Cas 1 : corrélation positive

Exemple 1 : 'Fertility' et 'N' avec 0.785580

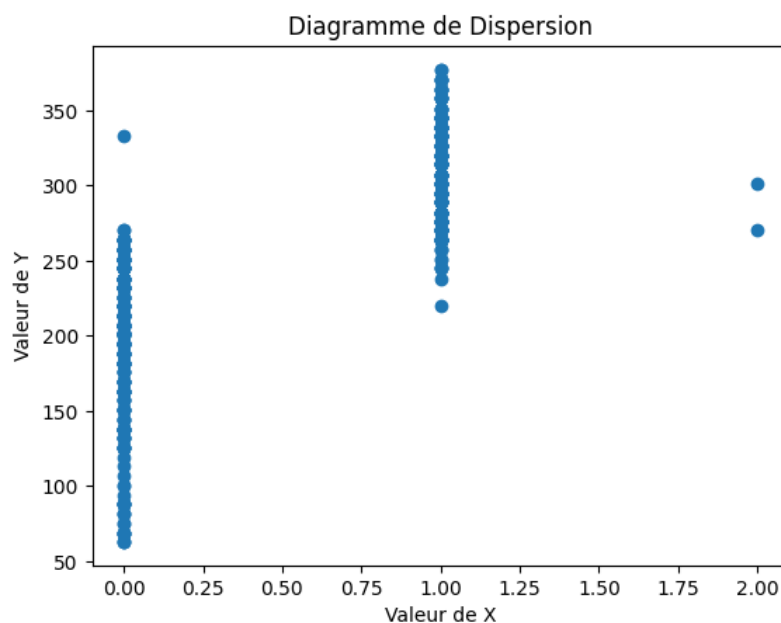


FIGURE 1.21 – Dispersion 'Fertility' et 'N'

Exemple 2 : 'OC' et 'OM' avec 1.002037

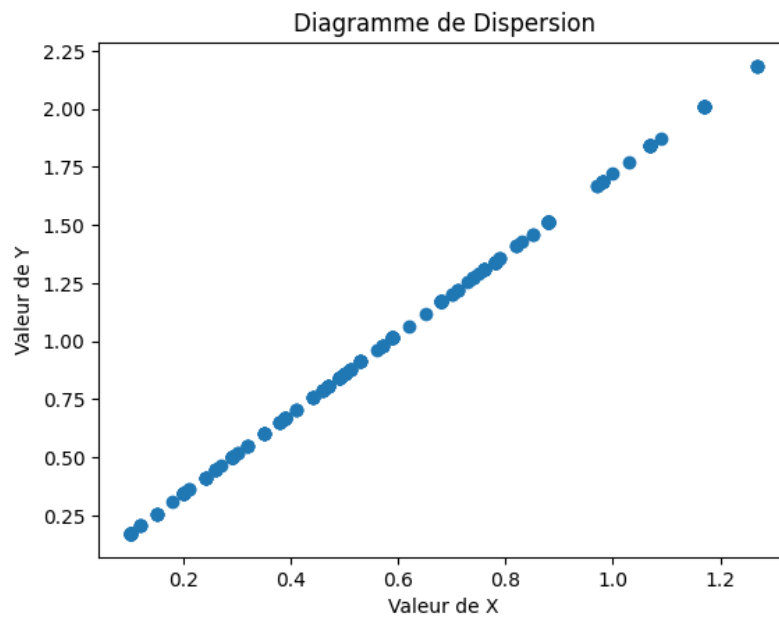


FIGURE 1.22 – Dispersion 'OC' et 'OM'

Cas 2 : corrélation négative

Exemple 1 : 'Ph' et 'Zn' avec -0.099292

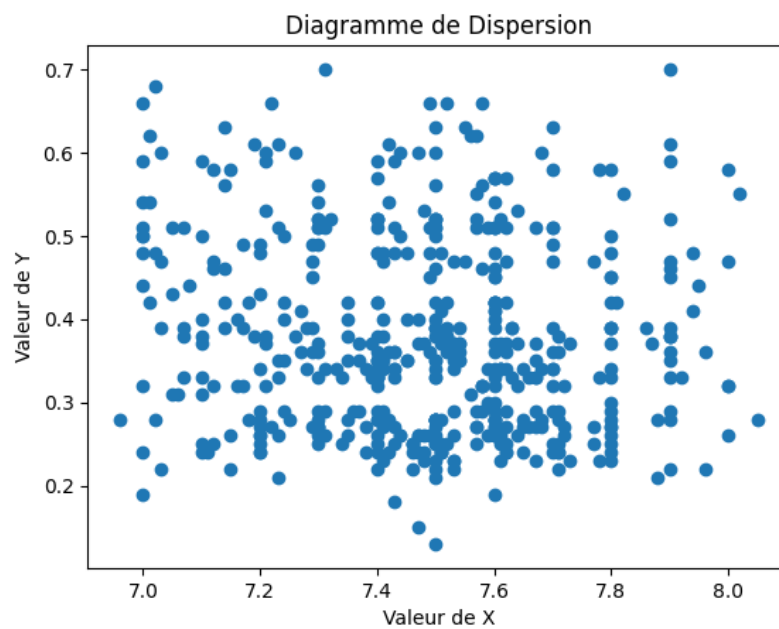


FIGURE 1.23 – Dispersion 'Ph' et 'Zn'

Exemple 2 : 'K' et 'EC' avec -0.139368

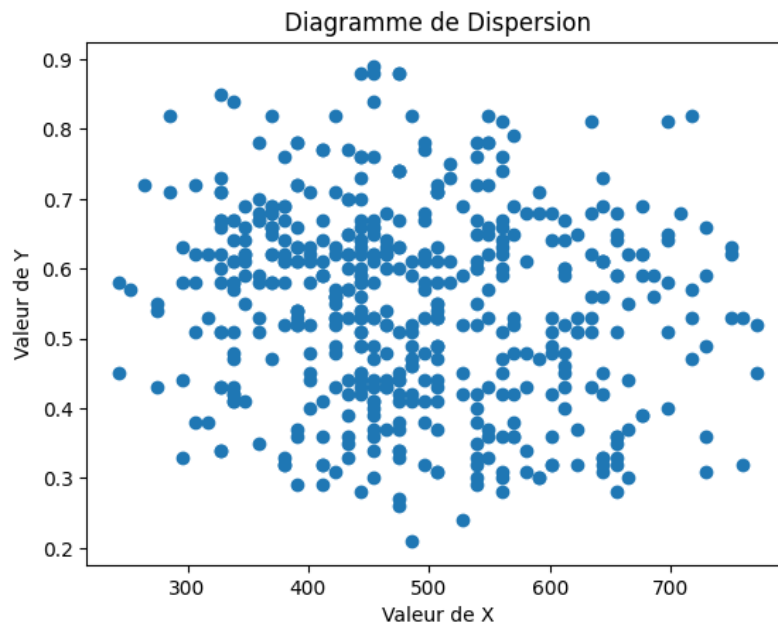


FIGURE 1.24 – Dispersion 'K' et 'EC'

Cas 3 : Absence de corrélation

Note : dans notre matrice des coefficients de corrélation on a pas trouvé un coefficient = 0 mais on peut prendre deux cas où leurs coefficients sont négligeables.

Exemple 1 : 'Mn' et 'N' avec -0.000499

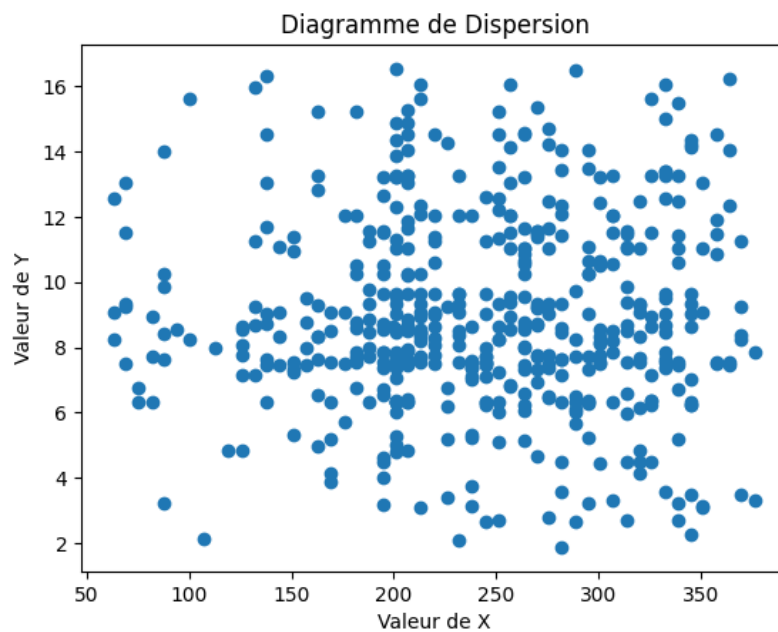


FIGURE 1.25 – Dispersion 'Mn' et 'N'

Exemple 2 : 'EC' et 'Zn' avec

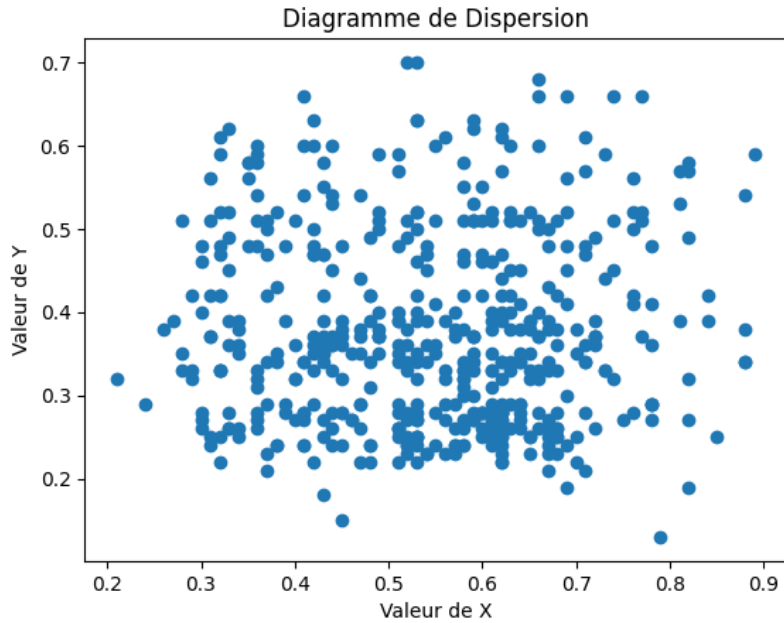


FIGURE 1.26 – Dispersion 'EC' et 'Zn'

2. Réduction des données (élimination des redondances)

L'élimination des redondances a pour objectif de réduire la dimensionnalité des données en supprimant des caractéristiques qui apportent peu d'informations distinctes ou qui sont fortement corrélées.

Dans le processus de réduction des données horizontales, nous éliminons les lignes dupliquées présentes dans le jeu de données, comme illustré dans l'image suivante du dataset après cette opération.

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
0	138.0	8.6	560.0	7.46	0.62	0.70	5.90	0.24	0.31	0.77	8.71	0.11	1.2040	0.0
1	245.0	8.1	560.0	7.31	0.63	0.78	11.60	0.29	0.43	0.57	7.73	0.74	1.3416	0.0
2	245.0	8.3	560.0	7.58	0.74	0.83	8.20	0.32	0.96	1.57	2.67	0.59	1.4276	0.0
3	107.0	8.3	612.0	7.43	0.43	0.75	7.20	0.18	0.95	0.89	2.16	0.74	1.2900	0.0
4	201.0	5.3	507.0	7.60	0.53	0.39	10.86	0.32	9.32	0.69	9.03	0.69	0.6708	0.0
5	151.0	5.9	581.0	7.60	0.48	0.88	4.52	0.42	4.05	1.06	5.30	0.29	1.5136	0.0
6	75.0	7.9	327.0	7.80	0.66	0.10	7.24	0.50	7.32	0.87	6.32	0.40	0.1720	0.0
7	232.0	5.9	275.0	7.70	0.43	0.68	5.43	0.47	6.52	1.52	8.05	0.32	1.1696	0.0
8	207.0	7.5	549.0	7.70	0.72	0.49	3.32	0.36	4.56	1.96	13.02	0.67	0.8428	0.0
9	264.0	6.4	496.0	7.90	0.38	0.78	3.92	0.52	8.06	1.32	14.52	0.46	1.3416	1.0

FIGURE 1.27 – Dataset après réduction H

Pendant la réduction verticale des données, nous éliminons l'un des deux attributs ayant le coefficient de corrélation maximal. Le choix de l'attribut à supprimer est basé sur la somme minimale de ses coefficients de corrélation avec les autres attributs. Dans notre cas les deux attributs ayant le coefficient de corrélation maximal sont : 'OC' et 'OM', l'attribut

supprimé est 'OM'.

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	Fertility
0	138.0	8.6	560.0	7.46	0.62	0.70	5.90	0.24	0.31	0.77	8.71	0.11	0.0
1	245.0	8.1	560.0	7.31	0.63	0.78	11.60	0.29	0.43	0.57	7.73	0.74	0.0
2	245.0	8.3	560.0	7.58	0.74	0.83	8.20	0.32	0.96	1.57	2.67	0.59	0.0
3	107.0	8.3	612.0	7.43	0.43	0.75	7.20	0.18	0.95	0.89	2.16	0.74	0.0
4	201.0	5.3	507.0	7.60	0.53	0.39	10.86	0.32	9.32	0.69	9.03	0.69	0.0
5	151.0	5.9	581.0	7.60	0.48	0.88	4.52	0.42	4.05	1.06	5.30	0.29	0.0
6	75.0	7.9	327.0	7.80	0.66	0.10	7.24	0.50	7.32	0.87	6.32	0.40	0.0
7	232.0	5.9	275.0	7.70	0.43	0.68	5.43	0.47	6.52	1.52	8.05	0.32	0.0
8	207.0	7.5	549.0	7.70	0.72	0.49	3.32	0.36	4.56	1.96	13.02	0.67	0.0
9	264.0	6.4	496.0	7.90	0.38	0.78	3.92	0.52	8.06	1.32	14.52	0.46	1.0

FIGURE 1.28 – Dataset après réduction V

3. Normalisation des données

La normalisation des données vise à mettre toutes les variables à la même échelle pour garantir que chaque variable contribue de manière équitable à l'apprentissage automatique. Nous utiliserons des techniques telles que la normalisation min-max et le z-score.

La normalisation de min-max se base sur la formule suivante :

$$X_{\text{normalisé}} = \frac{x - \min_{\text{old}}}{\max_{\text{old}} - \min_{\text{old}}} \cdot (\max_{\text{new}} - \min_{\text{new}}) + \min_{\text{new}}$$

où :

- \max_{old} est la valeur maximale observée dans l'ensemble de données original (old).
- \min_{old} est la valeur minimale observée dans l'ensemble de données original (old).
- \max_{new} est la valeur maximale souhaitée dans la plage normalisée (new).
- \min_{new} est la valeur minimale souhaitée dans la plage normalisée (new).

voici le résultat de cette normalisation sur notre dataset :

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	Fertility
0	0.238854	0.666667	0.600379	0.458716	0.602941	0.512821	0.368421	0.192982	0.009671	0.348718	0.410256	0.110000	0.000000
1	0.579618	0.587302	0.600379	0.321101	0.617647	0.581197	0.868421	0.280702	0.021277	0.246154	0.348718	0.740000	0.000000
2	0.579618	0.619048	0.600379	0.568807	0.779412	0.623932	0.570175	0.333333	0.072534	0.758974	0.071567	0.590000	0.000000
3	0.140127	0.619048	0.698864	0.431193	0.323529	0.555556	0.482456	0.087719	0.071567	0.410256	0.071567	0.690000	0.000000
4	0.439490	0.142857	0.500000	0.587156	0.470588	0.247863	0.803509	0.333333	0.881044	0.307692	0.410256	0.290000	0.000000
5	0.280255	0.238095	0.640152	0.587156	0.397059	0.666667	0.247368	0.508772	0.371373	0.497436	0.348718	0.400000	0.000000
6	0.038217	0.555556	0.159091	0.770642	0.661765	0.000000	0.485965	0.649123	0.687621	0.400000	0.348718	0.400000	0.000000
7	0.538217	0.238095	0.060606	0.678899	0.323529	0.495726	0.327193	0.596491	0.610251	0.733333	0.410256	0.320000	0.000000
8	0.458599	0.492063	0.579545	0.678899	0.750000	0.333333	0.142105	0.403509	0.420696	0.958974	0.758974	0.670000	0.000000
9	0.640127	0.317460	0.479167	0.862385	0.250000	0.581197	0.194737	0.684211	0.759188	0.630769	0.805000	0.460000	1.000000

FIGURE 1.29 – Normalisation min-max

La normalisation de z-score se base sur la formule suivante :

$$Z = \frac{X - \mu}{\sigma}$$

où Z est le Z-score, X est la variable, μ est la moyenne, et σ est l'écart-type.

voici le résultat de ce type de normalisation sur notre dataset :

	0	1	2	3	4	5	6	7	8	9
0	-1.475796	0.766263	0.667347	-0.084098	0.529371	0.432889	-0.254917	-1.177831	-1.829110	-0.41
1	0.038959	0.423098	0.667347	-0.719355	0.600884	0.695140	2.162866	-0.747557	-1.778561	-0.8
2	0.038959	0.560364	0.667347	0.424107	1.387527	0.859048	0.720680	-0.489393	-1.555305	1.4
3	-1.914651	0.560364	1.130178	-0.211150	-0.829377	0.596796	0.296507	-1.694160	-1.559517	-0.1
4	-0.583931	-1.498627	0.195614	0.508808	-0.114246	-0.583337	1.848979	-0.489393	1.966252	-0.5
5	-1.291760	-1.086829	0.854259	0.508808	-0.471812	1.022955	-0.840276	0.371155	-0.253677	0.2
6	-2.367662	0.285831	-1.406495	1.355816	0.815423	-1.533999	0.313474	1.059593	1.123774	-0.1
7	-0.145076	-1.086829	-1.869327	0.932312	-0.829377	0.367326	-0.454278	0.801429	0.786783	1.2
8	-0.498991	0.011299	0.569440	0.932312	1.244501	-0.255522	-1.349283	-0.145174	-0.038845	2.2
9	0.307935	-0.743664	0.097708	1.779321	-1.186942	0.695140	-1.094779	1.231703	1.435491	0.8

FIGURE 1.30 – Normalisation z-score

1.3 Données temporelles

1.3.1 Description Générale du Dataset 2

Le Dataset 2 offre une perspective détaillée de l'impact de la pandémie de COVID-19 dans plusieurs zones, identifiées par leur code postal. Composé de 337 instances, chaque ligne représente une période temporelle spécifique au sein d'une zone géographique donnée.

Vous trouverez ci-dessous une capture d'écran illustrant quelques-unes de ces instances du dataset 2 qui contient 337 lignes et 11 colonnes.

1	zcta	time_peri	population	Start date	end date	case count	test count	positive tests	case rate	test rate	positivity rate
2	95129	32	39741	10/11/2020	10/31/2020	22	2543	23	2.6	304.7	0.9
3	95129	43	39741	5/30/2021	6/19/2021		3315	14	1.1	397.2	0.4
4	95129	40	39741	3/28/2021	4/17/2021	34	4816	37	4.1	577.1	0.8
5	95129	55	39741	2/6/2022	2/26/2022	110	10194	175	13.2	1221.5	1.7
6	95129	44	39741	6/20/2021	7/10/2021	14	3033	17	1.7	363.4	0.6
7	95129	54	39741	1/16/2022	2/5/2022	624	13479	817	74.8	1615.1	6.1
8	95129	25	39741	5/17/2020	6/6/2020		762		0.4	91.3	0.4
9	95129	30	39741	8/30/2020	9/19/2020	20	1773	20	2.4	212.4	1.1
10	95129	31	39741	9/20/2020	10/10/2020	12	2120	12	1.4	254	0.6

FIGURE 1.31 – Quelques instances du Dataset 2

Voici une description globale des attributs du notre dataset :

- zcta : Code postal de la zone géographique concernée.
- time_period : Période temporelle spécifique pour chaque instance.
- population : La population totale de la zone géographique pendant la période donnée.
- Start date et End date : Les dates de début et de fin de la période pendant laquelle les données ont été recueillies.

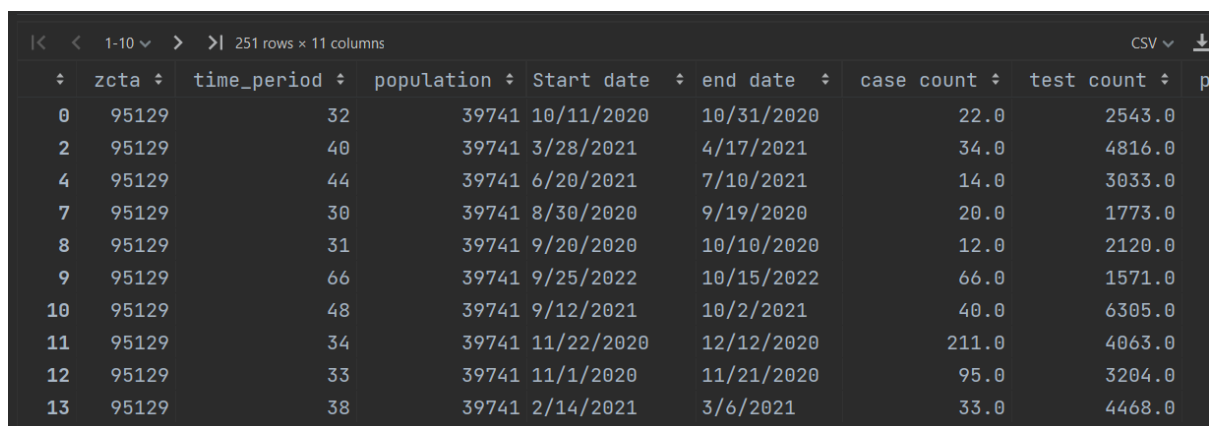
- case count : Le nombre total de cas de COVID-19 pendant la période.
- test count : Le nombre total de tests effectués pendant la période.
- positive tests : Le nombre de tests qui sont revenus positifs pour le COVID-19.
- case rate : Le taux de cas, indiqué en pourcentage.
- test rate : Le taux de tests effectués, également en pourcentage.
- positivity rate : Le taux de positivité des tests, exprimé en pourcentage.

1.3.2 Prétraitement

Suivant le même processus de prétraitement appliqué au Dataset 1, nous procéderons au prétraitement nécessaire du Dataset 2.

Nous avons opté pour une approche consistant à éliminer toutes les lignes contenant des valeurs manquantes ou aberrantes de notre ensemble de données.

Le dataset 2 après ce prétraitement est comme suit 251 lignes et 11 colonnes :



	zcta	time_period	population	Start date	end date	case count	test count
0	95129	32	39741	10/11/2020	10/31/2020	22.0	2543.0
2	95129	40	39741	3/28/2021	4/17/2021	34.0	4816.0
4	95129	44	39741	6/20/2021	7/10/2021	14.0	3033.0
7	95129	30	39741	8/30/2020	9/19/2020	20.0	1773.0
8	95129	31	39741	9/20/2020	10/10/2020	12.0	2120.0
9	95129	66	39741	9/25/2022	10/15/2022	66.0	1571.0
10	95129	48	39741	9/12/2021	10/2/2021	40.0	6305.0
11	95129	34	39741	11/22/2020	12/12/2020	211.0	4063.0
12	95129	33	39741	11/1/2020	11/21/2020	95.0	3204.0
13	95129	38	39741	2/14/2021	3/6/2021	33.0	4468.0

FIGURE 1.32 – Dataset 2

1.3.3 Visualisation

Afin de tirer les conclusions nécessaires, nous avons employé des représentations graphiques des données. Chaque type de graphique fournit des informations essentielles dans le contexte de l'extraction d'informations à partir du jeu de données.

Graphe 1 : La distribution du nombre total des cas confirmés et tests positifs par zones.

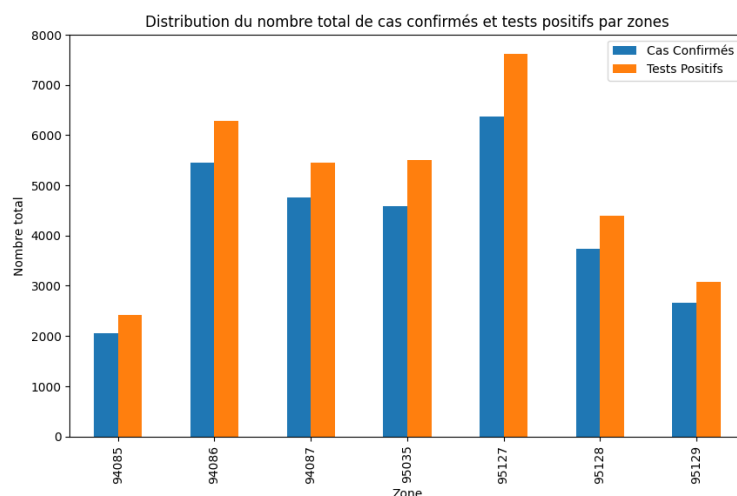


FIGURE 1.33 – Distribution

Observation : Le graphe de la distribution du nombre total des cas confirmés et des tests positifs par zones met en évidence des variations significatives entre les différentes zones. Certains secteurs présentent des niveaux élevés de cas confirmés et de tests positifs comme la zone 95127 suivie par la zone 94086, tandis que d'autres montrent une situation moins préoccupante comme la zone 94085. Cette diversité dans la répartition des données souligne l'importance d'analyser attentivement chaque zone.

NB : Le graphe peut être un peu différent à cause du choix des fonctions de prétraitement.

Graphe2 : L'évolution des tests COVID-19, les tests positifs et le nombre de cas évolue au fil du temps pour une zone choisit.

— La zone choisit est : 95129

L'évolution est hebdomadaire

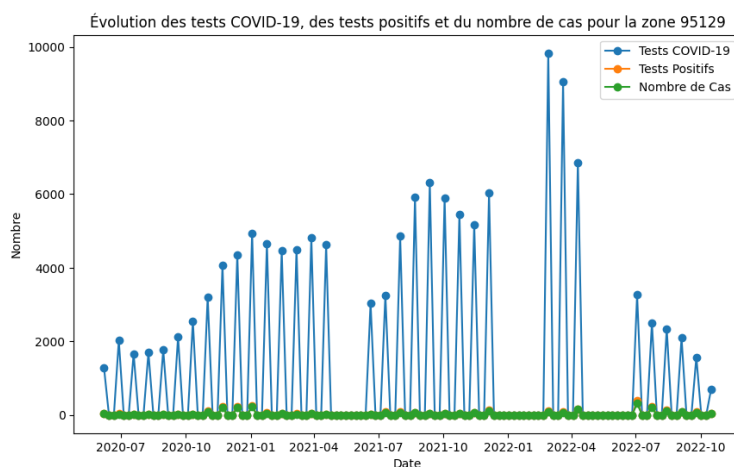


FIGURE 1.34 – Evolution hebdomadaire

Observation : Le graphique représentant l'évolution hebdomadaire des tests COVID-19, des tests positifs et du nombre de cas pour une zone spécifique offre une vue temporelle détaillée de la situation. On observe des fluctuations dans les trois catégories au fil des semaines, avec des moments de pic indiquant peut-être des périodes de propagation accrue du virus comme en 2022/04, mais toujours on a le nombre des tests effectués plus grand par rapport au nombre de tests positifs et au nombre de cas

- La zone choisit est : 95129
- L'évolution est mensuel

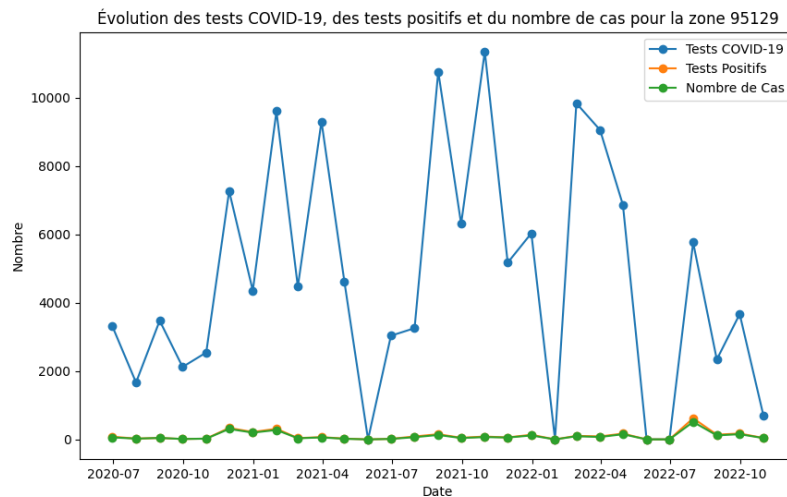


FIGURE 1.35 – Evolution mensuel

Observation : Dans le graphe mensuel pou la même zone 95129, on peut observer que le nombre de tests est toujours plus grands par rapport au deux autres attributs. Il y a des mois où le nombre de test diminue comme en 07/2021 et 01/2022.

- La zone choisit est : 95129
- L'évolution est annuel

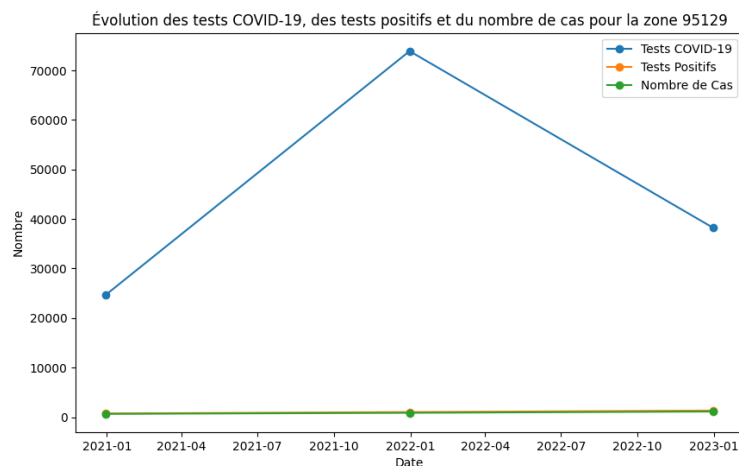


FIGURE 1.36 – Evolution annuel

Observation : Dans le graphe mensuel pou la même zone 95129, on ne peut pas bien observer les changements dans les troix attributs comme dans le graphe hebdomadaire et menseul mais il est claire que le nombre de tests effectuès est plus grands que les deux autres.

Graphe 3 : La distribution des cas covid positifs par zone et par année.

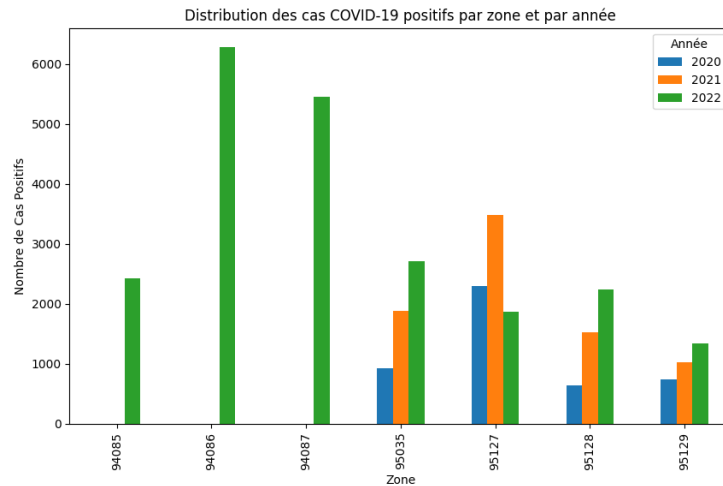


FIGURE 1.37 – Distribution cas covid positive

NB : Il existe des zones où l'année n'existe pas, on a choisit de les attribué l'année 2022.

Observation : En examinant la distribution des cas COVID-19 positifs par zone et par année, on remarque des variations marquées d'une année à l'autre. Certains secteurs, tels que la zone 95127, ont connu des pics significatifs en 2021. Ces variations temporelles soulignent l'importance d'une analyse approfondie pour mieux comprendre l'impact du COVID-19 sur différentes régions durant ces années.

Graphe 4 : La représentation du rapport entre la population et le nombre de tests effectués par zone.

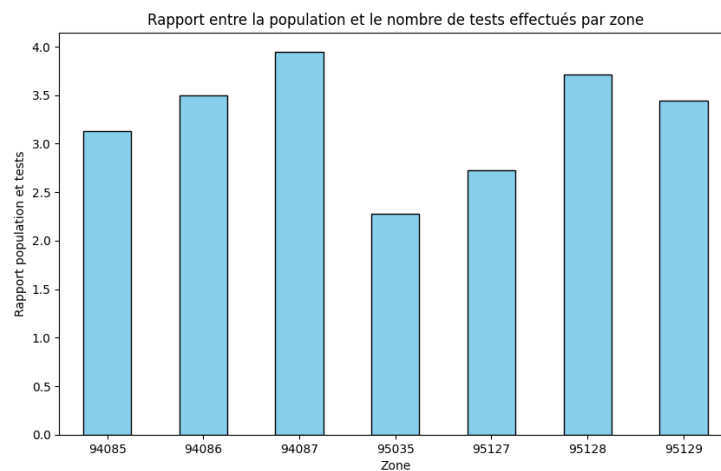


FIGURE 1.38 – Rapport population et test

Observation : Lors de l'analyse du graphique représentant le rapport entre la population et le nombre de tests effectués par zone, on remarque la barre de la zone 94087 est la plus grande cela indique que cette zone a le nombre de tests le plus élevé par rapport à sa population. À l'inverse, la zone 95035 a le rapport le plus bas entre la population et le nombre de tests effectués.

Graphe 5 : La distribution des testes positives par zone.

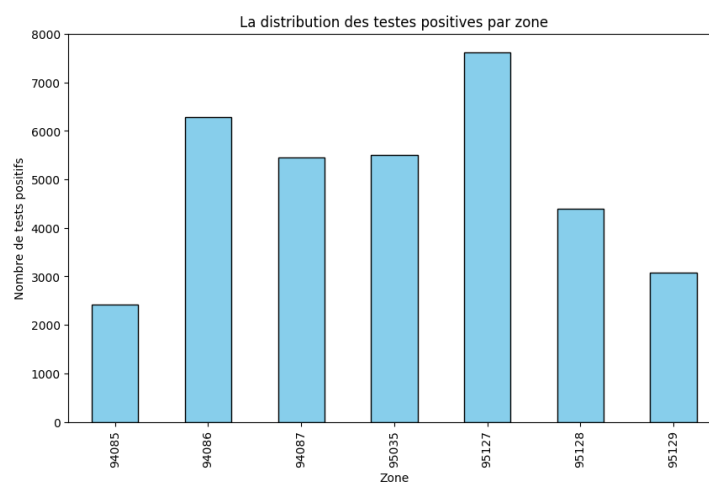


FIGURE 1.39 – Distribution testes positives

Observation : Après observation de graphe on peut déduire les zones les plus fortement impactées par le Coronavirus, par exemple les 5 premiers zones sont :

Zone 1 = 95127 , Zone 2 = 94086 , Zone 3 = 95035 , Zone 4 = 94087 , Zone 5 = 95128

Graphe 6 : Le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone, pour une période donnée .

Pour la visualisation des graphes on choisit les périodes ayant le plus grand nombre de zones possible.

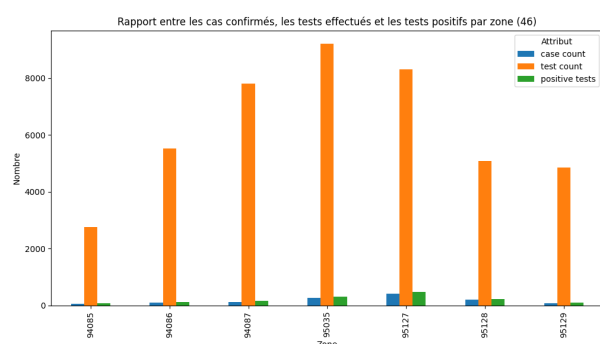


FIGURE 1.40 – Rapport p=46

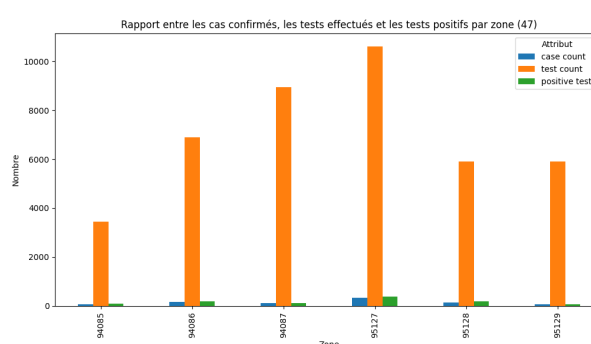


FIGURE 1.41 – Rapport p=47

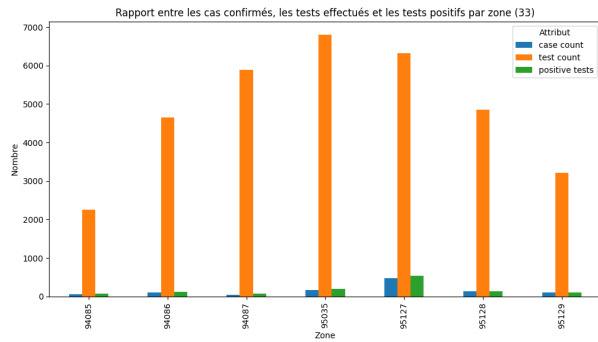


FIGURE 1.42 – Rapport p=33

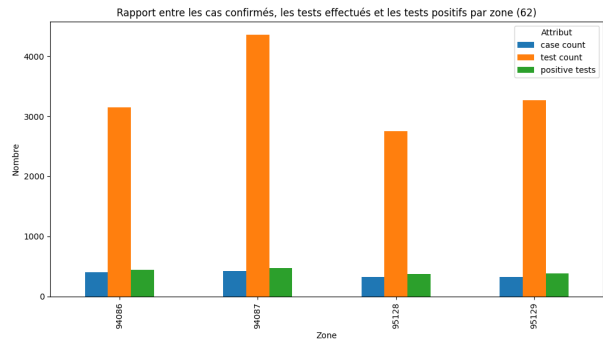


FIGURE 1.43 – Rapport p=62

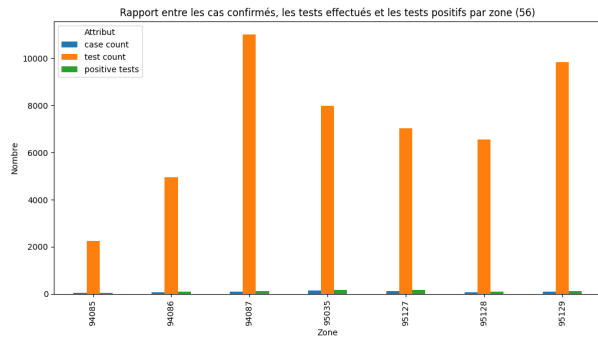


FIGURE 1.44 – Rapport p=56

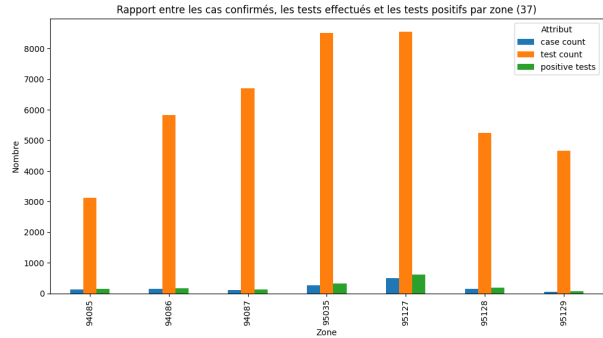


FIGURE 1.45 – Rapport p=37

Observation : En analysant les graphes du rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone avec des périodes différentes, on observe des variations marquées dans les attributs mentionnés pour différentes zones et périodes. Il est intéressant de noter que, presque systématiquement, le nombre de tests effectués demeure significativement élevé pour différentes zones et périodes. Cette constance souligne l'effort continu de réaliser un nombre substantiel de tests, indépendamment des fluctuations dans le nombre de cas confirmés. Nous avons également noté la présence de périodes où toutes les zones ne sont pas représentées, comme la période 62 représentée dans le graphe ci-dessus.

Chapitre 2

Extraction de motifs fréquents, règles d'associations et corrélations

2.1 Introduction

L'exploration des motifs fréquents, des règles d'associations et des corrélations au sein d'un dataset transactionnel ou relationnel offre des perspectives cruciales pour la prise de décisions dans divers domaines tels que le marketing, l'environnement et l'agriculture. Dans ce chapitre, nous nous concentrons sur l'analyse et l'extraction de motifs fréquents, de règles d'association et de corrélations à partir du dataset 3. Notre objectif est de mettre en évidence les relations entre les attributs liés au climat, au sol, à la végétation et à l'utilisation d'engrais, afin de fournir des informations essentielles pour la gestion des ressources environnementales et agricoles.

2.2 Discrétisation des Données : Approches Equal-Frequency et Equal-Width

La discrétisation des données en data mining consiste à transformer des variables continues en variables discrètes, en subdivisant la plage de valeurs continues en intervalles. Dans le cadre de notre étude, nous devons implémenter les deux méthodes suivantes :

- En classes d'effectifs égaux (equal-frequency)
- En classes d'amplitudes égales (equal-width).

2.2.1 Méthode 1 : Discrétisation en classes d'effectifs égaux

- Diviser les N valeurs possibles en Q quantiles (Q à définir).
- La position du i ème Quantile est égale à $\text{Position} = N * i / Q$.
- Toutes les valeurs appartenant à l'intervalle $[\text{Quantile } Q_i, Q_{i+1}[$ sont représentées par une même catégorie $0 \leq i \leq Q$.

2.2.2 Méthode 2 : Discrétisation en classes d'amplitudes égales

- Définir ou calculer le nombre d'intervalle k à utiliser.
- La largeur de chaque intervalle de valeurs est égale à : $(\text{ValeurMax} - \text{ValeurMin}) / k$.
- Toutes les valeurs appartenant à un même intervalle sont représentées par une même catégorie.

L'implémentation des deux méthodes a été réalisée dans notre code. Nous avons structuré notre approche de manière à ce que le choix de l'attribut à discrétiser soit paramétrable. Nous avons introduit un paramètre de fonction appelé `index_attribut`, qui indique le numéro de colonne de l'attribut que nous souhaitons discrétiser dans notre dataset.

NB : Avant d'effectuer la discrétisation des données, un prétraitement a été réalisé, consistant à éliminer les valeurs manquantes dans l'ensemble des données.

2.3 Extraction des motifs fréquents puis les règles d'association en utilisant l'algorithme Apriori

Dans cette section, nous explorons en détail l'utilisation de l'algorithme Apriori pour extraire des motifs fréquents et déduire des règles d'association significatives à partir de notre datasets.

2.3.1 Extraction des motifs fréquents

Description

L'algorithme Apriori, développé par Agrawal et Srikant, est un algorithme efficace pour extraire des motifs fréquents à partir de données transactionnelles. Son fonctionnement repose sur une approche itérative appelée recherche par niveaux, visant à identifier des ensembles d'items fréquents et à explorer de manière systématique des motifs significatifs dans les données.

Identification des itemsets individuels fréquents : Initialement, l'algorithme identifie tous les itemsets individuels fréquents dans le Dataset.

Génération d'itemsets de taille supérieure : Ensuite, il génère des itemsets de taille supérieure en utilisant ces itemsets fréquents précédemment découverts et ainsi de suite...

Pseudo-code

```

 $C_k$  : Ensemble d'items candidats de taille k
 $L_k$  : Ensemble d'items fréquents de taille k
 $L_1 = \{\text{items fréquents}\}$ 
for  $k = 1 ; L_k \neq \emptyset ; k++$  do
     $C_{k+1}$  = candidats générés à partir de  $L_k$ 
    for chaque transaction t dans dataset do
        Incréments le compte de tous les candidats dans  $C_{k+1}$  qui sont présents
        dans  $t$ 
    end
     $L_{k+1}$  = candidats dans  $C_{k+1}$  avec un support supérieur ou égal à  $min\_support$ 
end
return  $\bigcup_k L_k$ 

```

NB : Le support d'un itemset est définie par :

$$\text{Support_Item} = \frac{\text{Nombre de transaction dans lesquelles apparaît l'item}}{\text{Nombre total de transactions}}$$

Sachant que le support minimum est un paramètre de l'algorithme d'Apriori.

Application de l'Algorithme Apriori au Dataset 3

Description Générale du Dataset 3 :

Le Dataset 3 utilisé dans cette étude regroupe des informations essentielles liées aux conditions environnementales dans le contexte agricole. Il comprend les paramètres tels que Temperature , Humidity , Rainfall , soil, corp, et Fertilizer.

1	Tempera	Humidity	Rainfall	Soil	Crop	Fertilizer
2	24,87	82,84	295,61	Clayey	rice	DAP
3	28,69	96,65	178,96	laterite	Coconut	Good NPK
4	20,27	81,64	270,44	silty clay	rice	MOP
5	25,07	95,02	192,9	sandy	Coconut	Urea
6	25,04	95,9	174,8	coastal	Coconut	Urea
7	20,82	84,13	230,22	clay loam	rice	Urea
8	25,95	93,41	172,05	alluvial	Coconut	Urea
9	26,49	80,16	242,86	Clayey	rice	DAP
10	25,01	95,59	165,81	coastal	Coconut	Urea
11	21,87	80,19	224,56	silty clay	rice	Urea

FIGURE 2.1 – Quelques instances du Dataset 3

Prétraitement des Données : Avant d'appliquer l'algorithme Apriori au Dataset 3, un processus de prétraitement des données a été mis en œuvre pour garantir la qualité et la pertinence des résultats.

1. Traitement des valeurs manquantes : nous avons choisi la méthode de suppression des valeurs manquantes en raison de sa simplicité et de son efficacité dans le contexte de notre dataset.
2. Discrétisation des attributs : Nous avons discrétisé tous les attributs continus de dataset, notamment l'Humidity, Temperature et Rainfall, en trois catégories : Low , Medium et high, en utilisant la méthode basée sur des classes d'amplitudes égales.

Application de l'Algorithme : Une fois les données prétraitées, Nous avons appliqué l'algorithme Apriori sur dataset après avoir défini une transaction comme une ligne et les items comme les valeurs uniques de chaque attribut. En fixant le support minimal à 0.4, nous avons extrait les K-item sets fréquents.

Les résultats sont les suivants :

```
Les K-item Set Pour MinSup= 0.4
Coconut ----- > Avec support = 0.5152542372881356
low_Humidity ----- > Avec support = 0.5050847457627119
medium_Temperature ----- > Avec support = 0.4542372881355932
medium_Rainfall ----- > Avec support = 0.4440677966101695
rice ----- > Avec support = 0.4847457627118644
low_Humidity - rice ----- > Avec support = 0.4542372881355932
```

FIGURE 2.2 – Motifs Fréquents Extraits avec un Support Minimum de 0.5

Analyse des résultats : Les résultats de l'application de l'algorithme Apriori ont mis en lumière des motifs fréquents significatifs au sein de notre ensemble de données agricoles.

- **Item Coconut (Support : 0.515)** : La culture du cocotier est prédominante, observée dans plus de 51% des transactions, soulignant ainsi son rôle significatif dans nos pratiques agricoles.
- **Item Low Humidity (Support : 0.505)** Les conditions de basse humidité sont fréquentes dans plus de 50% des transactions, révélant une prévalence marquée de ces conditions dans notre contexte agricole.
- **Item Medium Temperature (Support : 0.454)** Les conditions de température moyenne sont présentes dans près de 45% des transactions, démontrant ainsi une diversité de pratiques agricoles liées à la température.
- **Item Medium Rainfall (Support : > 44%)** Les niveaux moyens de précipitations sont fréquents, dépassant 44%, ce qui pourrait influencer le choix des cultures adaptées à ces conditions spécifiques de précipitations.
- **Item Rice (Support : 0.485)** La culture du riz est présente dans près de 49% des transactions, soulignant son importance notable dans notre ensemble de données agricoles.
- **Item Low Humidity and Rice (Support : 0.454)** La combinaison de basse humidité et de la culture du riz est fréquente dans près de 45% des transactions, soulignant une corrélation significative entre ces deux conditions.

2.3.2 Extraction des règles d'associations et corrélations

Règles d'Associations

Une règle d'association est une implication de la forme :

$$X \rightarrow Y \text{ [support, confiance]}$$

Où X et Y représentent des ensembles d'items, le support mesure la fréquence d'apparition de l'ensemble $X \cup Y$ dans le dataset et la Confiance indique la probabilité conditionnelle de Y étant présent lorsque X est présent. Ces règles nous permettent d'identifier des modèles significatifs dans dataset.

Métriques de Qualité

1. **Confidence** : La confiance mesure la fiabilité de la règle. Une confiance élevée indique une forte dépendance entre les deux ensembles. Cette mesure est particulièrement utile pour filtrer les règles et ne retenir que celles qui sont hautement fiables.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

2. **Lift** : Le lift compare la probabilité de co-occurrence observée de X et Y avec ce à quoi on pourrait s'attendre s'ils étaient indépendants. Un lift supérieur à 1 indique une association plus forte que celle attendue par hasard.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) \times \text{Support}(Y)}$$

3. **Conviction** : La conviction évalue la confiance en l'absence de Y étant conditionnelle à X . Une conviction élevée indique que la règle est robuste même en l'absence de Y . Elle est utile pour identifier les règles où la présence de X a un impact significatif sur l'absence de Y . Une conviction élevée indique une forte dépendance entre X et $\neg Y$ (l'absence de Y).

$$\text{Conviction}(X \rightarrow Y) = \frac{1 - \text{Support}(Y)}{1 - \text{Confidence}(X \rightarrow Y)}$$

4. **Cosine** : Le cosinus mesure la similarité entre les ensembles X et Y , il indique dans quelle mesure les deux ensembles partagent des éléments communs par rapport à leur taille respective. Un cosinus proche de 1 indique une forte similarité, ce qui peut être utile pour identifier des patterns similaires dans le dataset. C'est particulièrement pertinent lorsque la fréquence des items est importante, mais leur co-occurrence n'est pas fréquente.

$$\text{Cosine}(X, Y) = \frac{\text{Support}(X \cap Y)}{\sqrt{\text{Support}(X) \times \text{Support}(Y)}}$$

Extraction des Règles d'Associations à Partir du Dataset 3

Suite à l'application de l'algorithme Apriori, conformément à la description dans la section 2.3.1, un ensemble de motifs fréquents a été obtenu. Nous procédons ensuite à la génération de règles d'association en se basant sur ces motifs, en fixant la confiance minimale à 0.4. Les résultats sont présentés dans la Figure 2.3

```
Les regle d'association Pour confiance min = 0.4
-----> Coconut ( Avec Confiance = 0.5152542372881356 )
-----> low_Humidity ( Avec Confiance = 0.5050847457627119 )
-----> medium_Temperature ( Avec Confiance = 0.4542372881355932 )
-----> medium_Rainfall ( Avec Confiance = 0.4440677966101695 )
-----> rice ( Avec Confiance = 0.4847457627118644 )
-----> low_Humidity AND rice ( Avec Confiance = 0.4542372881355932 )
low_Humidity -----> rice ( Avec Confiance = 0.8993288590604026 )
rice -----> low_Humidity ( Avec Confiance = 0.937062937062937 )
```

FIGURE 2.3 – Règles d'Association Générées avec une Confiance Minimale de 0.5

Analyse des résultats : En examinant les règles d'association extraites par l'algorithme Apriori, nous identifions des associations significatives, telles que :

- La règle associant la présence du riz dans toutes les transactions ($\rightarrow \text{rice}$) avec une confiance de 48.5%. Cette observation suggère l'existence de conditions particulières ou de facteurs favorables à la croissance du riz.
- De plus, les deux règles extraites, ($\text{low_Humidity} \rightarrow \text{rice}$ et $\text{rice} \rightarrow \text{low_Humidity}$), présentent toutes deux une confiance dépassant 89%, suggérant ainsi une association significative dans les deux directions.

2.3.3 Expérimentation

Dans le but d'évaluer l'influence des deux paramètres, à savoir la confiance minimale et le support minimal, des expérimentations seront réalisées. Les résultats obtenus sont présentés ci-dessous :

- **Pour $\text{minSup} = 0.1$ et $\text{minConf} = 0.1$:** nous identifions 86 k-items fréquents et déduisons 449 règles associatives, dont plus de 39 % ont une valeur de lift proche de 1. Cela indique que la probabilité d'occurrence simultanée des éléments dans l'association est similaire à ce à quoi on pourrait s'attendre au hasard. Parmi les règles on observe :
 1. $\rightarrow \text{medium_Humidity}$
 $\{ \text{confiance} = 0.169, \text{lift} = 1, \text{Conviction} = 1, \text{Cosinus} = 2.429 \}$
 2. $\text{medium_Temperature} \rightarrow \text{high_Rainfall AND low_Humidity}$
 $\{ \text{confiance} = 0.239, \text{lift} = 1.174, \text{Conviction} = 1.046, \text{Cosinus} = 3.290 \}$
 3. $\text{rice} \rightarrow \text{Clayey AND high_Rainfall}$
 $\{ \text{confiance} = 0.224, \text{lift} = 2.063, \text{Conviction} = 1.148, \text{Cosinus} = 4.361 \}$
- **Pour $\text{minSup} = 0.1$ et $\text{minConf} = 0.5$:** Dans ce cas, le nombre de règles va diminuer, et nous trouvons 82 règles, avec 11 % des valeur lift proche de 1, les regles qu'on peut trouve :
 1. $\rightarrow \text{low_Humidity}$
 $\{ \text{confiance} = 0.505, \text{lift} = 1, \text{Conviction} = 1.0, \text{Cosinus} = 1.407 \}$
 2. $\text{Clayey} \rightarrow \text{high_Rainfall AND rice}$
 $\{ \text{confiance} = 0.593, \text{lift} = 2.91, \text{Conviction} = 1.95, \text{Cosinus} = 5.183 \}$
- **Pour $\text{minSup} = 0.1$ et $\text{minConf} = 0.9$:** Quand la confiance minimale est établie à 0.9, nous découvrons des règles de haute qualité (17 règles), illustrées par un lift supérieur à 1, indiquant ainsi une forte cohérence dans les associations identifiées.
 1. $\text{silty clay} \rightarrow \text{low_Humidity}$
 $\{ \text{confiance} = 0.985, \text{lift} = 1.949, \text{Conviction} = 32.161, \text{Cosinus} = 2.998 \}$
 2. $\text{high_Humidity} \rightarrow \text{Coconut}$
 $\{ \text{confiance} = 0.958, \text{lift} = 1.86, \text{Conviction} = 11.63, \text{Cosinus} = 2.442 \}$
- **Pour $\text{minSup} = 0.3$ et $\text{minConf} = 0.1$:** En augmentant le seuil de minSup à 0.3, seuls les motifs présents dans au moins 30 % des transactions sont considérés comme fréquents, entraînant ainsi une réduction significative du nombre de motifs fréquents à 28. En fixant la confiance minimale à 0.1, on identifie 21 règles, parmi lesquelles plus du 61 % ont une valeur de lift proche de 1.
 1. $\rightarrow \text{low_Rainfall}$
 $\{ \text{confiance} = 0.35, \text{lift} = 1, \text{Conviction} = 1, \text{Cosinus} = 1.68 \}$
- **Pour $\text{minSup} = 0.3$ et $\text{minConf} = 0.5$:** On obtient 10 règles, dont 20 % ont une valeur de lift proche de 1.
 1. $\rightarrow \text{low_Humidity}$
 $\{ \text{confiance} = 0.5, \text{lift} = 1, \text{Conviction} = 1.0, \text{Cosinus} = 1.4 \}$
 2. $\text{Coconut} \rightarrow \text{high_Temperature}$
 $\{ \text{confiance} = 0.59, \text{lift} = 1.69, \text{Conviction} = 1.69, \text{Cosinus} = 2.35 \}$
- **Pour $\text{minSup} = 0.3$ et $\text{minConf} = 0.9$:** On a obtenu seulement deux règles qui sont :
 1. $\text{high_Humidity} \rightarrow \text{Coconut}$
 $\{ \text{confiance} = 0.958, \text{lift} = 1.85, \text{Conviction} = 1.85, \text{Cosinus} = 2.44 \}$
 2. $\text{rice} \rightarrow \text{low_Humidity}$
 $\{ \text{confiance} = 0.937, \text{lift} = 1.855, \text{Conviction} = 1.855, \text{Cosinus} = 2.02 \}$

Analyse des résultats :

En analysant ces résultats, il est possible de constater que l'ajustement de minSup et minConf a un impact significatif sur le nombre et la spécificité des règles générées. Une valeur plus élevée de minSup et minConf conduit à des règles plus restrictives, tandis qu'une valeur plus basse permet une plus grande diversité mais peut inclure des règles moins fiables.

Conclusion :

le choix d'un minSup bas (0.1) associé à une confiance min (0.5) élevée peut être le meilleur choix dans notre cas.

Conclusion générale

En conclusion globale de cette partie du projet, l'exploration approfondie et le prétraitement des deux datasets, statique et temporel, ont jeté les bases nécessaires pour une analyse avancée dans le deuxième chapitre du projet. La première étape a permis d'acquérir une compréhension approfondie des propriétés du sol, de l'évolution du COVID-19 au fil du temps, et de la distribution des tests et cas positifs par zone.

La deuxième étape, axée sur l'extraction de motifs fréquents, de règles d'association et de corrélations, vise à révéler les relations complexes entre les attributs liés au climat, au sol, à la végétation et à l'utilisation d'engrais. Cette analyse avancée offre des perspectives essentielles pour la prise de décisions informées dans les domaines de la gestion des ressources environnementales et agricoles.

L'ensemble du projet illustre la puissance du data mining dans l'extraction d'informations exploitables à partir de datasets complexes et variés. Les résultats obtenus ouvrent des perspectives prometteuses pour la compréhension approfondie des relations entre différents attributs. Ce travail sert de base pour des analyses plus avancées et des recommandations éclairées dans des contextes environnementaux, agricoles et épidémiologiques.

Bibliographie

[1] Data Mining Concepts and Techniques Jiawei Han Jian Pei Hanghang Tong 4th Edition

[2] Serie Exercice 3 Pré-traitement des données

[3] Cours Data mining - Chapter 3 : inspired from the following textbook