République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche

Université des Sciences et de la Technologie Houari Boumediene

FACULTÉ D'INFORMATIQUE
Département Intelligence Artificielle et Science des Données

Filière : Informatique

Spécialité : Systèmes Informatiques Intelligents

RAPPORT Projet 1

Exploitation des données et Extraction des règles d'associations

---

Binôme :

DJELLAB Nesrine

MEKKI Feriel

Sommaire

Table des Tableaux

Table des Figures

Table des Graphes

Introduction

L'évolution rapide des systèmes de recommandation a profondément influencé divers secteurs, améliorant l'expérience utilisateur et facilitant la prise de décisions éclairées. Notre projet, intitulé **"Exploitation des données et Extraction des règles d'associations"**, s'inscrit dans cette tendance en apportant l'efficacité des systèmes de recommandation à un domaine essentiel : l'agriculture.

Dans ce contexte spécifique, nous nous concentrons sur la recommandation de types de graines et d'engrais, des éléments cruciaux pour optimiser les pratiques agricoles. Notre démarche se divise en trois parties bien définies, chacune contribuant de manière significative à la réalisation de notre objectif global.

La première section s'attèle à l'analyse et au prétraitement exhaustifs des trois datasets qui représentent respectivement des données statiques, temporelles, et climatiques. L'objectif est d'assurer la fiabilité et la pertinence des résultats pour les étapes suivantes du projet, notamment la recommandation dans le domaine agricole.

La deuxième partie se concentre sur la création de visualisations significatives à partir du "dataset 2", mettant en lumière l'évolution des cas COVID-19 aux États-Unis de 2019 à 2023. Cette analyse visuelle approfondie offre des perspectives claires sur la propagation de la maladie.

La troisième partie se consacre à la création d'un système de recommandation en utilisant l'algorithme Apriori, appliqué au "dataset 3". Notre objectif est d'explorer les motifs fréquents, les règles d'association, et les corrélations entre les données liées au climat, au sol, à la végétation, et à l'utilisation d'engrais. Cette approche vise à fournir des recommandations pertinentes pour la gestion efficace des ressources agricoles et environnementales.

En résumé, notre projet aspire à démontrer la valeur ajoutée des systèmes de recommandation dans le domaine de l'agriculture. En suivant une approche méthodique, de l'analyse des données aux recommandations pratiques, nous cherchons à maximiser l'efficacité des pratiques agricoles grâce à l'intelligence des systèmes de recommandation.

Partie I: Analyse et prétraitement des données

Introduction

La première étape cruciale dans tout processus d'exploration de données et de mise en œuvre d'algorithmes d'apprentissage automatique réside dans l'analyse approfondie et le prétraitement des données. Cette phase constitue le fondement sur lequel repose la qualité des résultats obtenus ultérieurement. L'objectif principal de la Partie I est d'explorer les données disponibles, d'identifier les caractéristiques pertinentes et d'appliquer des techniques de prétraitement pour garantir la fiabilité et la qualité des données utilisées dans le cadre de l'étude.

Au cours de cette partie, nous aborderons différentes méthodes d'analyse exploratoire des données afin de comprendre la nature et la distribution des variables. Nous nous pencherons également sur la détection et la gestion des valeurs aberrantes, la manipulation des données manquantes et la normalisation des caractéristiques pour garantir une homogénéité dans le jeu de données.

L'importance de cette phase ne peut être sous-estimée, car des données de qualité médiocre peuvent compromettre la performance des modèles prédictifs et analytiques. En fournissant une introduction approfondie à l'analyse et au prétraitement des données, cette partie jettera les bases nécessaires pour la compréhension et l'interprétation des résultats obtenus tout au long de l'étude.

Prétraitement des Données pour l'Analyse de la Fertilité du Sol

Introduction

L'analyse de la fertilité du sol est cruciale pour le secteur agricole, car elle permet de comprendre les composants du sol qui influent sur la croissance des cultures. Dans cette étude, nous avons entrepris une analyse de data mining en utilisant un ensemble de données statiques sur la fertilité du sol. Ce rapport détaille la démarche suivie, du prétraitement initial à la compréhension des caractéristiques du dataset.

Description du dataset

Aperçu des Caractéristiques du Dataset

Le dataset comprend 885 entrées, chacune associée à 14 colonnes représentant différentes caractéristiques du sol. Les colonnes comprennent des informations telles que les niveaux de nutriments tels que l'azote (N), le phosphore (P), le potassium (K), le pH, la conductivité électrique (EC), la matière organique (OM), et d'autres éléments tels que le zinc (Zn), le fer (Fe), le cuivre (Cu), le manganèse (Mn), le bore (B), et la fertilité du sol.

Types de Données

Les types de données dans le dataset comprennent principalement des entiers (int64), des nombres à virgule flottante (float64), et un objet. La colonne 'P' est de type objet, ce qui nécessitera une transformation ultérieure si une analyse quantitative est envisagée.

Analyse approfondie du jeu de données sur la fertilité du sol pour l'optimisation agricole

À noter: cette analyse servira de base lors du traitement des outliers en utilisant la méthode du binning.
Le jeu de données examiné constitue une ressource cruciale pour les praticiens et les chercheurs en agriculture, car il encapsule des informations essentielles sur la fertilité du sol. Comprendre les détails complexes des propriétés du sol est impératif pour optimiser les pratiques agricoles et assurer une croissance végétale durable. Notre approche a impliqué une exploration approfondie de différents articles scientifiques pour déterminer les intervalles réels des nutriments du sol, formant la base de la description du jeu de données.

Le jeu de données se compose de plusieurs attributs, chacun représentant une propriété spécifique du sol. Voici une interprétation concise de chaque attribut :

N (Azote) : Essentiel à la croissance des plantes, le contenu en azote dans le sol est fourni dans l'intervalle [100, 500]. [1]
P (Phosphore) : Autre nutriment vital pour le développement des plantes, le contenu en phosphore se situe dans la plage [0, 80]. [2]
K (Potassium) : Reflétant le contenu en potassium, essentiel à la santé globale des plantes, l'intervalle est [160, 880]. [3][4][5]
pH : La mesure de l'acidité ou de l'alcalinité du sol, cruciale pour la disponibilité des nutriments, se situe dans la plage [3, 9]. [6]

CE (Conductivité électrique) : Indiquant la salinité ou la fertilité du sol, la conductivité électrique est représentée dans l'intervalle [0,11, 0,57]. [7]
OC (Carbone organique) : La quantité de carbone organique, composant clé de la matière organique du sol, est fournie dans l'intervalle [0, 14]. [8]
S (Soufre) : Essentiel pour les plantes, le contenu en soufre est dans l'intervalle [0, 18].[9]
Zn (Zinc) : Micronutriment essentiel à la croissance des plantes, le contenu en zinc se situe dans [0,12, 2,17].[10]
Fe (Fer) : Autre micronutriment important, le contenu en fer est représenté dans l'intervalle [0,2, 55].[11]
Cu (Cuivre) : Le micronutriment cuivre est représenté dans l'intervalle [0, 3.0].[12]
Mn (Manganèse) : Essentiel pour divers processus métaboliques des plantes, le contenu en manganèse se situe dans [0,1, 13].[12]
B (Bore) : Le contenu en bore, micro nutriment important pour le développement des plantes, est dans l'intervalle [0,04, 7,40].[13]
MO (Matière organique) : Indique la teneur globale en matière organique dans le sol, cruciale pour la structure et la fertilité du sol, dans l'intervalle [0, 20].[14]
Fertilité : Représente une mesure globale de la fertilité du sol, probablement calculée en fonction de divers niveaux de nutriments et d'autres propriétés du sol.

Comprendre les intervalles des nutriments du sol est essentiel pour adapter les pratiques agricoles afin d'optimiser la croissance des plantes. Ce jeu de données constitue une base complète pour des pré-traitements ultérieurs.

Analyse des Valeurs Manquantes

La gestion des valeurs manquantes est une étape cruciale dans le prétraitement des données, garantissant la qualité et la fiabilité des analyses ultérieures.
Nous avons effectué une analyse approfondie des valeurs manquantes dans chaque colonne du dataset. Voici un résumé des résultats :

Attribut	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
Nbr	0	2	0	0	0	1	0	0	0	1	0	0	0	0
%	0	0.23	0	0	0	0.11	0	0	0	0.11	0	0	0	0

La majorité des colonnes présentent un pourcentage de valeurs manquantes nul ou négligeable. Cependant, il est essentiel de traiter les quelques valeurs manquantes dans les colonnes 'P', 'OC', et 'Cu' avant de poursuivre les analyses. Les approches courantes telles que l'imputation ou la suppression des lignes concernées peuvent être envisagées pour garantir l'intégrité des données. Cette analyse préliminaire des valeurs manquantes jettera les bases d'une exploration plus approfondie des relations entre les caractéristiques du sol et la fertilité.

Analyse des Caractéristiques des Attributs du Dataset de Fertilité du Sol

Analyse des tendances centrales

L'analyse des caractéristiques des attributs est une étape essentielle dans la compréhension approfondie du dataset de fertilité du sol. Cette analyse vise à déterminer les mesures de tendance centrale telles que la moyenne, la médiane, et le mode pour chaque attribut, fournissant ainsi des insights significatifs sur la distribution des données.

Voici les résultats de l'analyse pour chaque attribut du dataset :

Attribut	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
μ	246.99	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
Q2	257	7	475	7.5	0.55	0.59	6.64	0.36	3.56	0.93	8.34	0.41	1.01	1
Mode	207	X	444	7.5	0.53 0.62	0.88	4.22 5.13	0.28	6.32	1.25	7.54	0.34	1.51	1
Symmetrie	N	N	N	O	N	N	N	N	N	N	N	N	N	N

Tableau 1.1: Tendances centrales du dataset 1

Les résultats montrent que la plupart des attributs du dataset ne suivent pas une distribution symétrique. Cela suggère une asymétrie dans la répartition des données, ce qui peut être utile pour identifier des tendances ou des schémas spécifiques. L'absence de symétrie dans la distribution des données souligne la diversité des caractéristiques du sol dans l'ensemble du dataset. Ces résultats serviront de base pour des analyses plus approfondies visant à comprendre les relations entre les différentes caractéristiques du sol et la fertilité globale.

Analyse des quartiles

L'analyse des quartiles des attributs constitue une étape cruciale pour comprendre la répartition des données du dataset de fertilité du sol. Cette analyse fournit des informations sur la dispersion des valeurs, permettant ainsi d'identifier la variabilité et les tendances au sein des différentes caractéristiques du sol.

Voici les résultats de l'analyse des quartiles pour chaque attribut du dataset :

Attribut	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
Q0	6	10.1	11	0.9	0.1	0.1	0.64	0.07	0.21	0.09	0.11	0.06	0.172	0
Q1	201	4.8	412	7.35	0.43	0.38	4.7	0.28	2.05	0.63	6.221	0.27	0.6536	0
Q2	257	7	475	7.5	0.55	0.59	6.64	0.36	3.56	0.93	8.34	0.41	1.0148	1
Q3	307	8.1	581	7.63	0.64	0.78	8.75	0.47	6.32	1.25	11.47	0.61	1.3416	1
Q4	383	125	1560	11.15	0.95	24	31	42	44	3.02	31	2.82	41.28	2

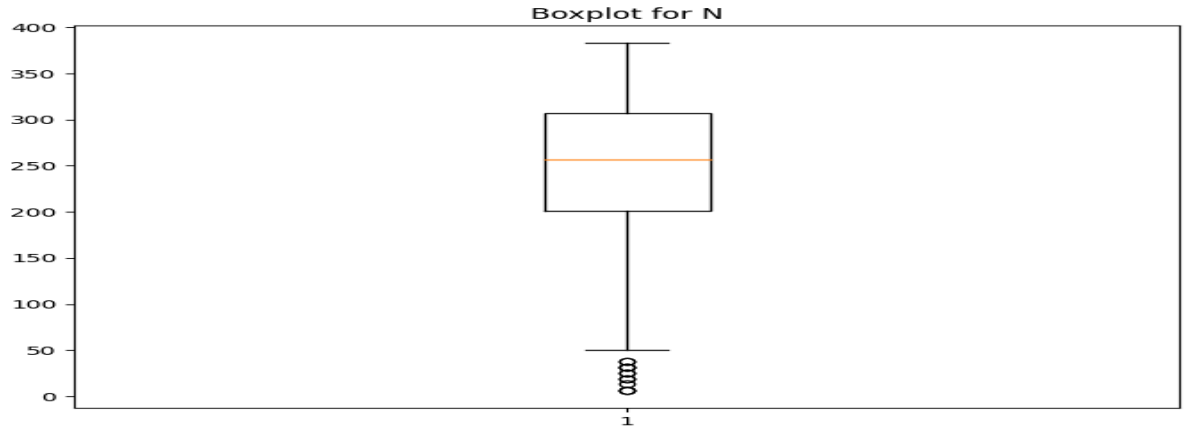
Tableau 1.2: Quartiles du dataset 1

L'analyse des quartiles offre une perspective détaillée sur la répartition des valeurs pour chaque attribut du dataset de fertilité du sol. Les quartiles Q0, Q1, Q2 (médiane), Q3, et Q4 fournissent des informations clés sur la dispersion des données, identifiant ainsi les valeurs extrêmes, la variabilité interquartile, et la médiane qui représente le point central de la distribution.

Analyse des Boxplots et Valeurs Aberrantes

L'analyse des boxplots et la détection des valeurs aberrantes sont des étapes essentielles pour comprendre la distribution des données et identifier les points qui s'écartent de la normale. Cette analyse nous permet d'observer la variabilité des différentes caractéristiques du sol.

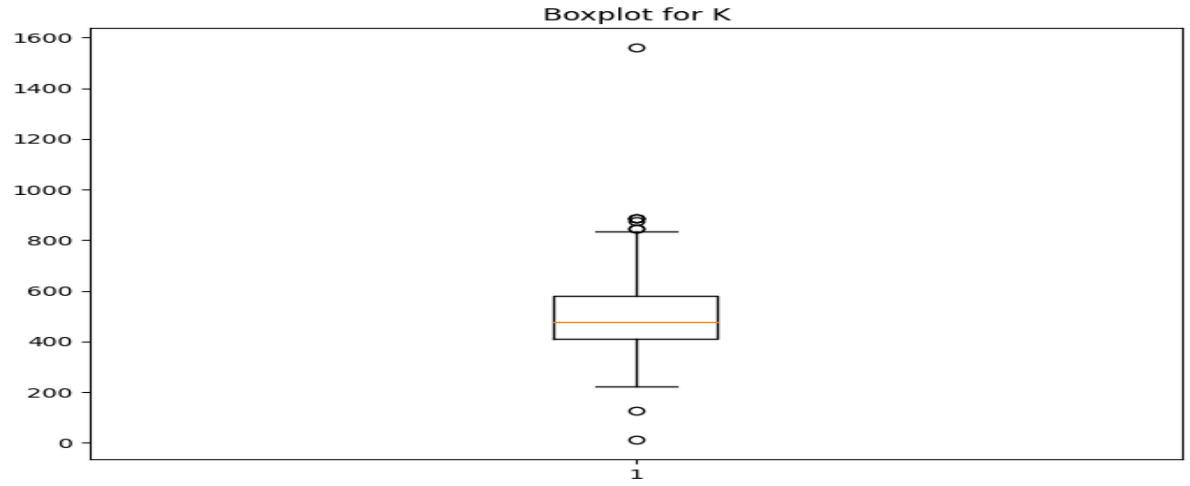
Attribut : N (Azote)



Graphe 1.1: Boxplot de N

Cet attribut présente des données entre les intervalles 0 et 400 comme vu lors de l'étude des quartiles.
Le boxplot pour l'attribut "N" révèle plusieurs valeurs aberrantes, notamment aux positions 38, 19, 25, 31, et 6. Ces valeurs sont en dehors des limites définies par le boxplot et peuvent nécessiter une attention particulière lors de l'analyse. Les valeurs anormales peuvent indiquer des erreurs de mesure ou des conditions exceptionnelles dans les données.

Attribut : K (Potassium)

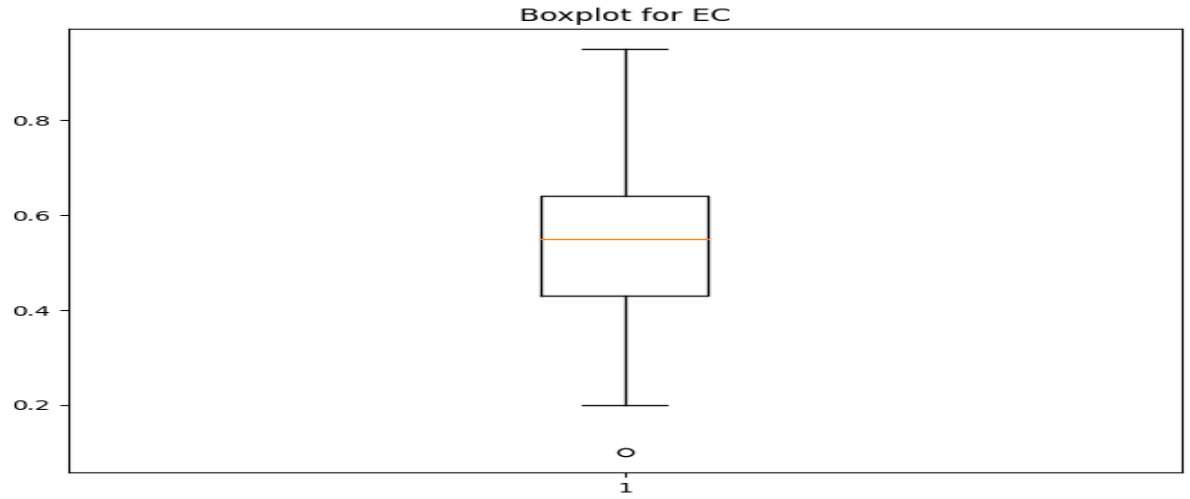


Graphe 1.2: Boxplot de K

Cet attribut présente des données entre les intervalles 0 et 1600 comme vu lors de l'étude des quartiles.
Pour l'attribut "K", plusieurs valeurs aberrantes sont également identifiées, notamment aux positions 127, 11, 887, 876, 845, et 1560. Ces valeurs, dépassant les limites du boxplot,

méritent une investigation approfondie pour comprendre la raison de leur écart par rapport à la distribution générale.

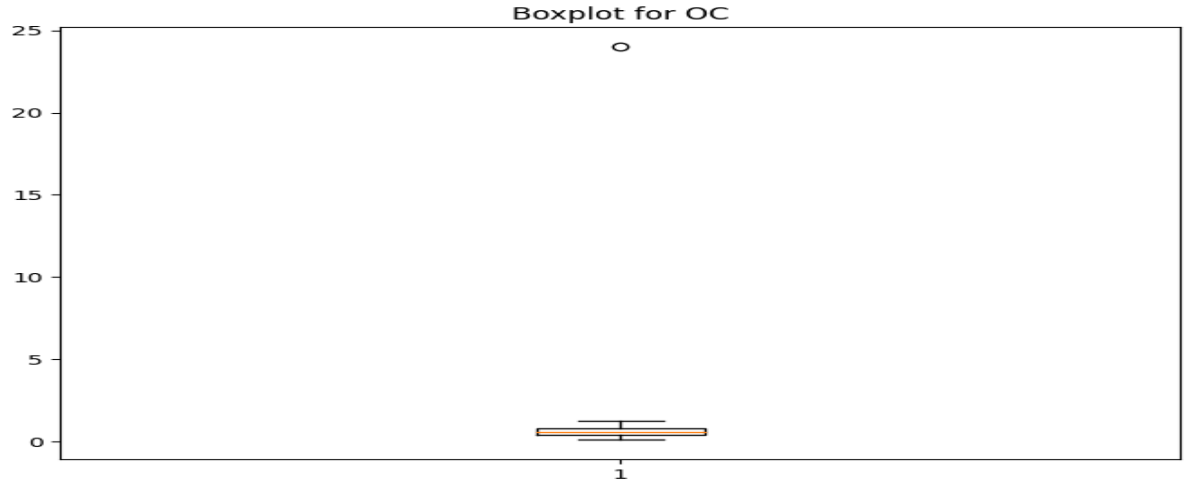
Attribut : EC (Conductivité Électrique)



Graphe 1.3: Boxplot de EC

Cet attribut présente des données entre les intervalles 0 et 1 comme vu lors de l'étude des quartiles.
Une seule valeur aberrante est observée pour l'attribut "EC" à la position 0.1. Cette valeur pourrait indiquer une anomalie ou nécessiter une vérification supplémentaire pour assurer la qualité des données.

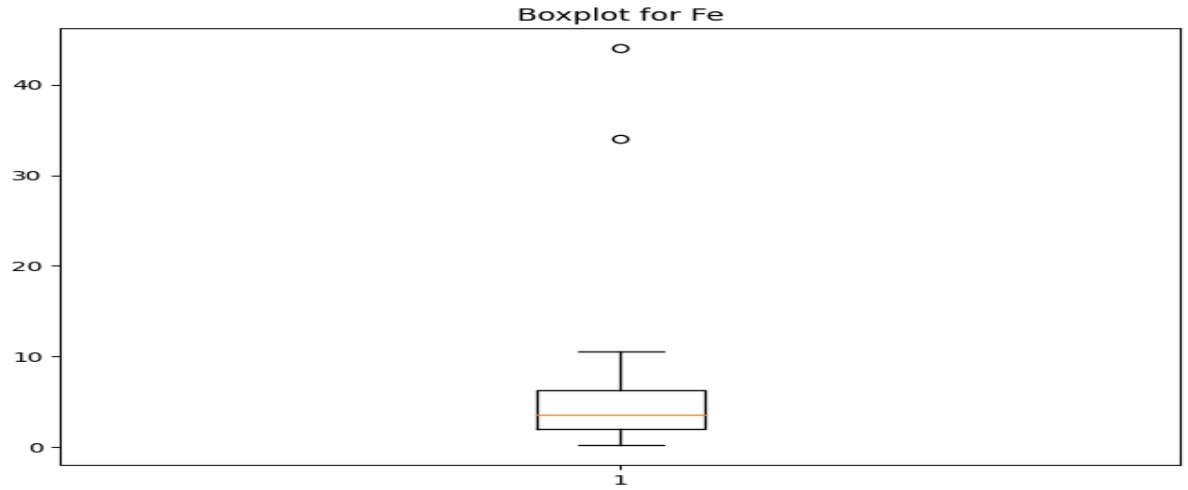
Attribut : OC (Carbone Organique)



Graphe 1.4: Boxplot de OC

Cet attribut présente des données entre les intervalles 0 et 25 comme vu lors de l'étude des quartiles.
L'attribut "OC" présente une valeur aberrante à la position 24.0. L'origine de cette valeur atypique doit être évaluée pour garantir l'intégrité des données.

Attribut : Fe (Fer)



Graphe 1.5: Boxplot de fe

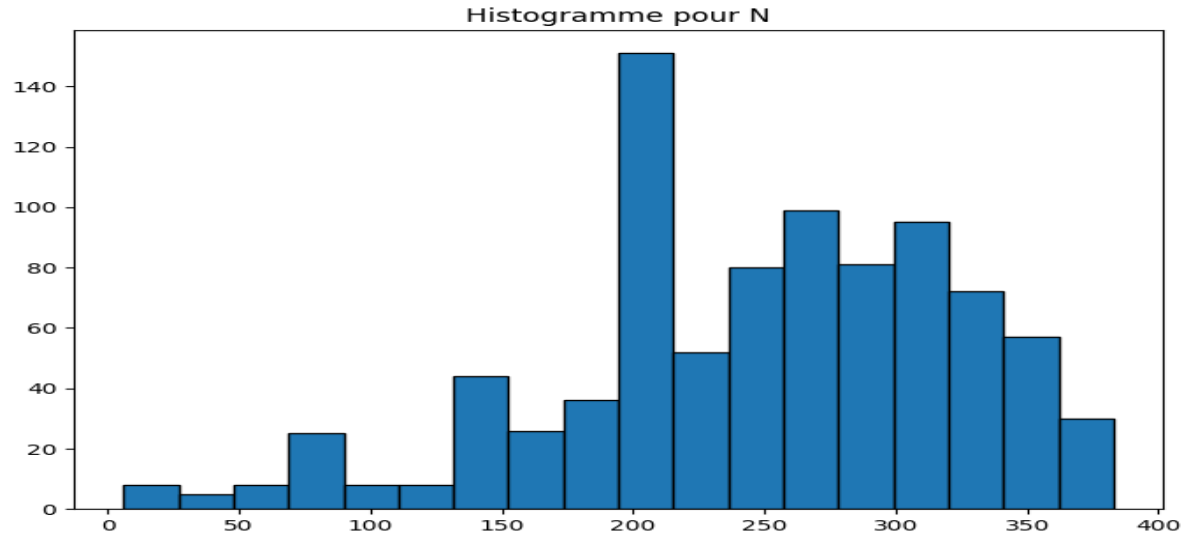
Cet attribut présente des données entre les intervalles 0 et 40 comme vu lors de l'étude des quartiles.
Pour l'attribut "Fe", des valeurs aberrantes sont observées aux positions 44.0 et 34.0. Ces points nécessitent une analyse plus approfondie pour déterminer s'ils sont liés à des erreurs de mesure ou à des conditions exceptionnelles.

Analyse des Histogrammes

Cette visualisation permet de mieux comprendre la répartition des valeurs et d'identifier d'éventuelles tendances, modes, ou anomalies dans les données.

Voici les histogrammes pour chaque attribut du dataset :

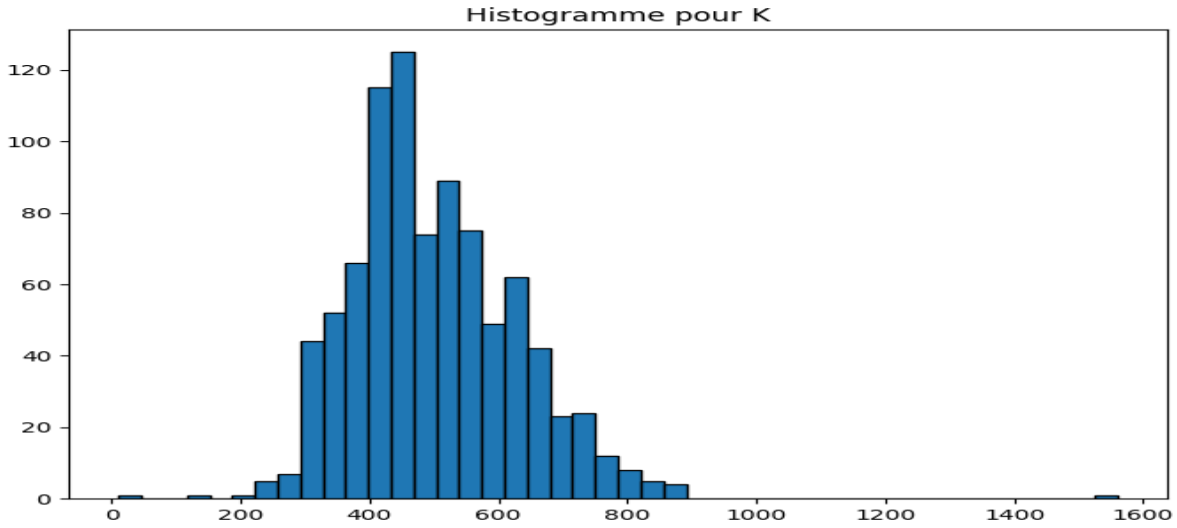
Attribut : N (Azote)



Graphe 1.6: Histogramme de N

L'histogramme pour l'attribut "N" montre une distribution relativement asymétrique avec une concentration plus élevée autour des valeurs centrales. Cependant, la présence de quelques valeurs élevées indique une asymétrie négative.

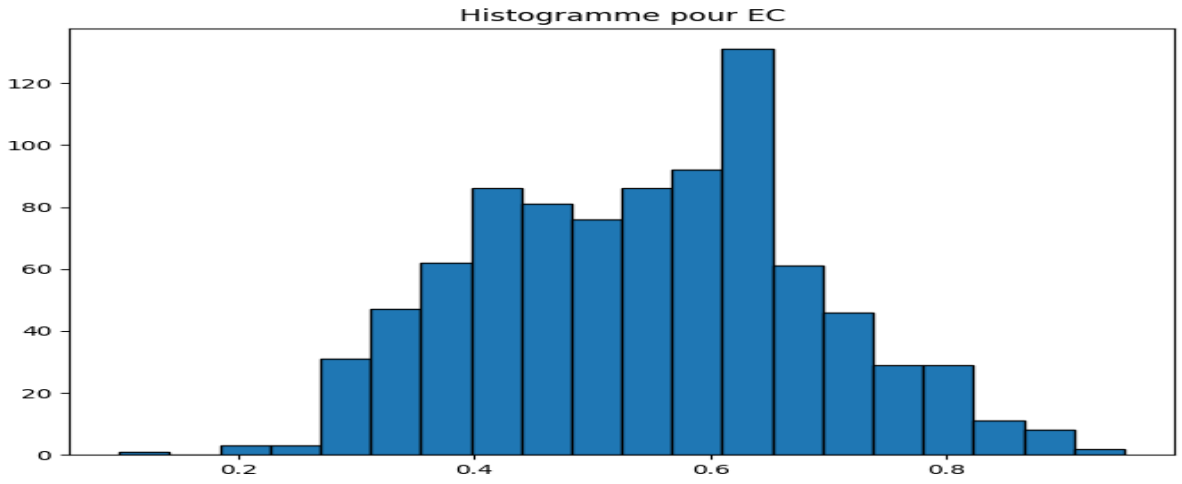
Attribut : K (Potassium)



Graphe 1.7: Histogramme pour K

L'histogramme pour l'attribut "K" indique une distribution asymétrique positive, bien que quelques valeurs élevées puissent influencer la forme de la distribution.

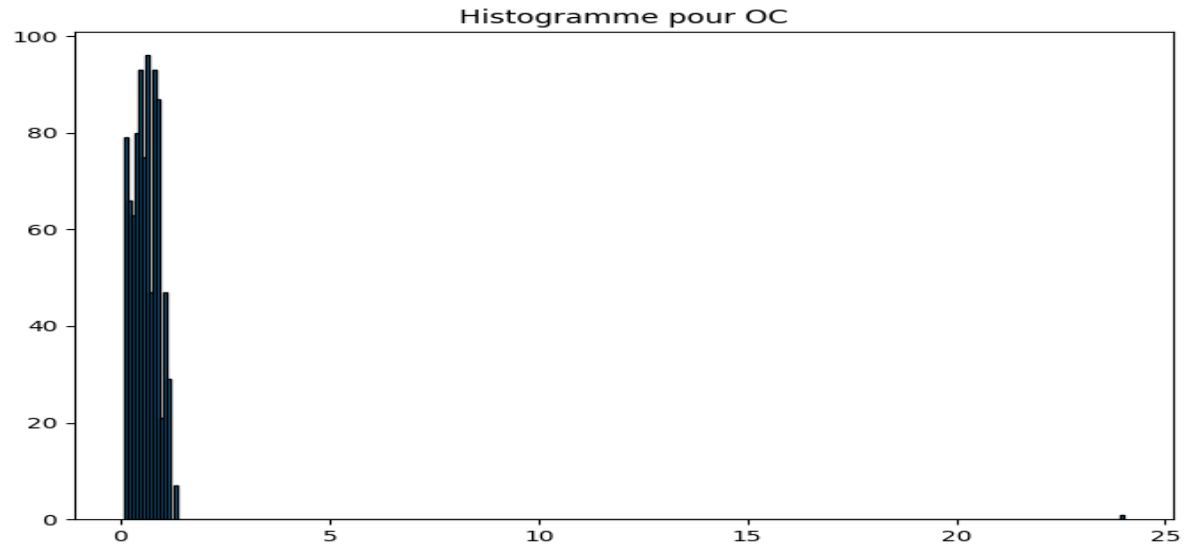
Attribut : EC (Conductivité Électrique)



Graphe 1.8: Histogramme pour EC

L'histogramme pour l'attribut "EC" révèle une distribution asymétrique positive, avec la majorité des valeurs concentrées du côté gauche de la moyenne.

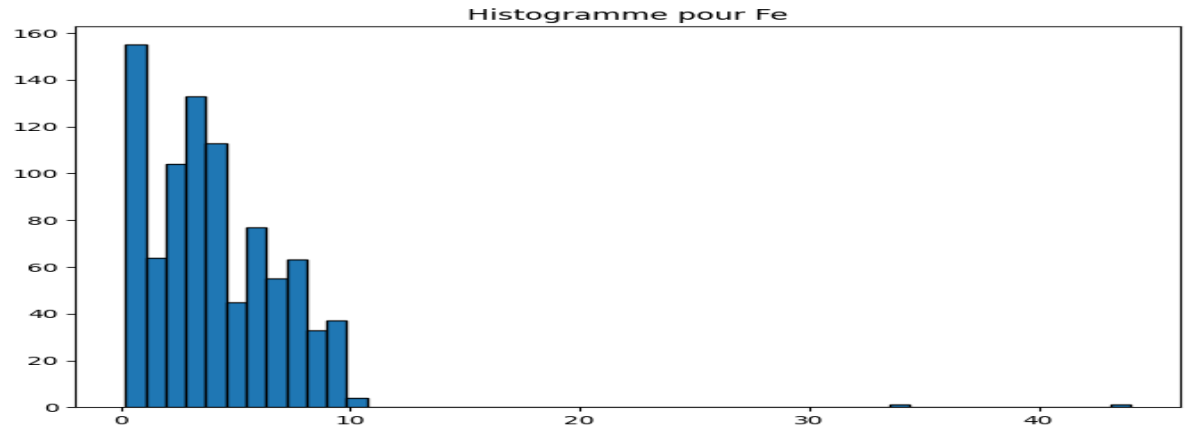
Attribut : OC (Carbone Organique)



Graphe 1.9: Histogramme pour OC

L'histogramme pour l'attribut "OC" montre une distribution asymétrique positive, indiquant une concentration plus élevée de valeurs inférieures à la moyenne.

Attribut : Fe (Fer)



Graphe 1.10: Histogramme pour FE

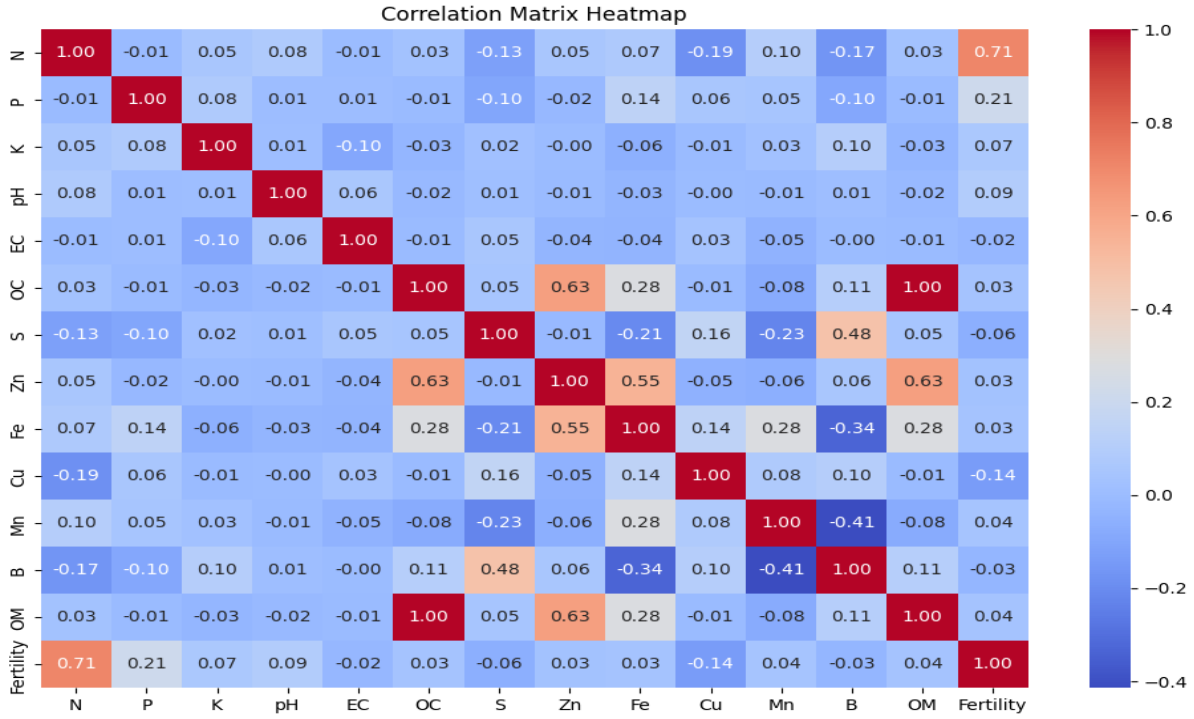
L'histogramme pour l'attribut "Fe" révèle une distribution asymétrique positive, indiquant une concentration plus élevée de valeurs inférieures à la moyenne.

Étude de la Corrélation entre les Caractéristiques du Sol

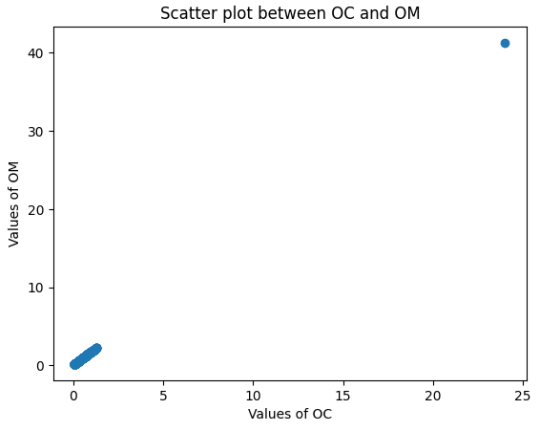
L'analyse de la corrélation entre les caractéristiques du sol est essentielle pour comprendre les relations entre différentes variables et identifier d'éventuelles dépendances. Cette section explore la corrélation entre les attributs de l'ensemble de données sur la fertilité du sol.

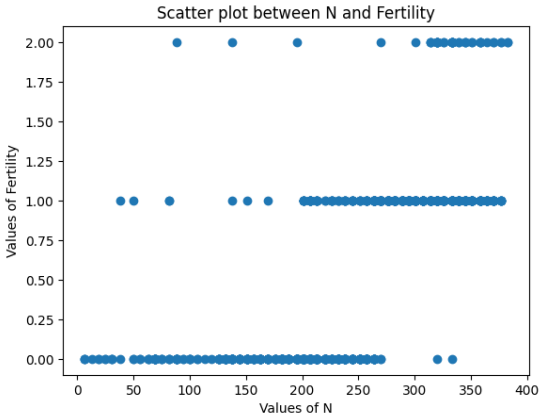
Nous avons utilisé la bibliothèque Seaborn pour créer une matrice de corrélation. Cette matrice met en évidence les relations linéaires entre les différentes caractéristiques du sol. Les valeurs de corrélation varient de -1 à 1, où 1 indique une corrélation positive forte, -1 une corrélation négative forte, et 0 l'absence de corrélation linéaire.

La matrice de corrélation ci-dessous présente les coefficients de corrélation entre chaque paire d'attributs.



Graphe 1.11: Matrice de corrélation

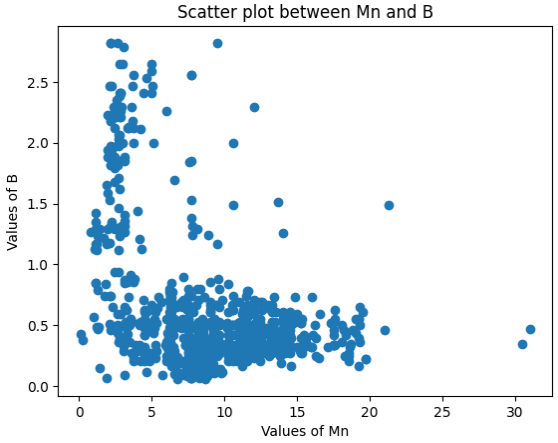


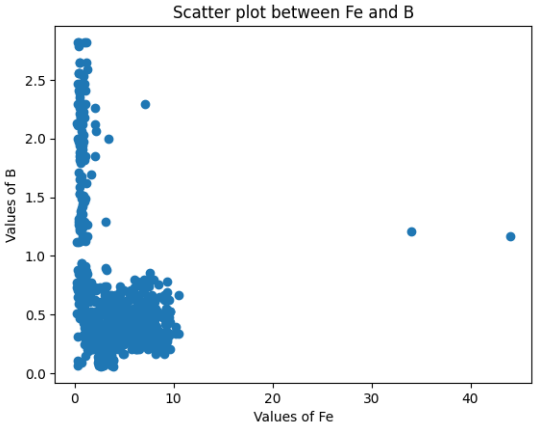


Graphe 1.12: Scatter plot entre OC et OM

Graphe 1.13: Scatter plot entre N et Fertility

On observe que les caractéristiques OM (Matière Organique) et OC (Carbone Organique) démontrent une corrélation positive très forte, atteignant une valeur de 1. De même, les variables N (Azote) et Fertilité présentent également une corrélation positive notable, évaluée à 0.71. Ce qui signifie que lorsque la valeur d'une caractéristique augmente, l'autre a tendance à augmenter également. Comme vu dans les scatter plots ci-dessous ().

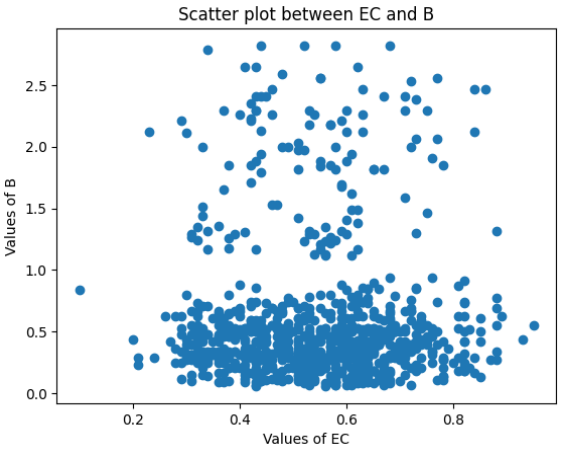




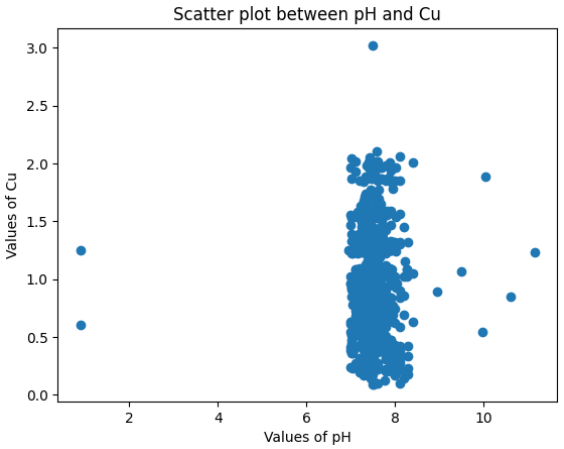
Graphe 1.12: Scatter plot entre MN et B

Graphe 1.13: Scatter plot entre Fe et B

En revanche, nous avons observé des corrélations négatives entre les caractéristiques B (Bore) et Mn (Manganèse), avec un coefficient de -0.41. De manière similaire, une corrélation négative de -0.34 a été mise en évidence entre les attributs Fe (Fer) et B (Bore). Suggérant que lorsque la valeur d'une caractéristique augmente, l'autre a tendance à diminuer. Comme le démontrent les scatter plots ci- dessous.



Graphe 1.14: Scatter plot entre EC et B



Graphe 1.15: Scatter plot entre Ph et Cu

Enfin, il est intéressant de noter que les attributs EC et B, ainsi que Cu et pH, ne présentent aucune corrélation, affichant un coefficient de 0. Cela suggère qu'ils ne sont pas du tout corrélés entre eux dans le contexte de notre analyse. Ce qui indique une absence de corrélation linéaire. Comme nous pouvons l'observer sur les scatter plots ci- contres.

Prétraitement du dataset

Le prétraitement du dataset est une étape essentielle dans le processus d'analyse de données, visant à traiter les valeurs de notre jeu de données de manière appropriée pour garantir l'intégrité et la qualité des données.

Traitement des valeurs manquantes

Dans cette partie, nous décrirons la méthode utilisée pour remplacer les valeurs manquantes dans le dataset de fertilité du sol.

Étape 1 : Identification des Valeurs Inappropriées
La première étape consiste à identifier les valeurs inappropriées dans chaque attribut. Dans notre approche, nous utilisons des expressions régulières pour déterminer si une valeur est de type entier, décimal, ou incompatible avec ces deux formats. Les valeurs incompatibles sont remplacées par NaN (Not a Number).

Étape 2 : Remplacement des NaN par la Moyenne par Classe
Après avoir identifié les valeurs manquantes, nous remplaçons ces NaN par la moyenne des valeurs de l'attribut correspondant, en tenant compte de la classe de fertilité. Cette approche permet de conserver la variabilité des données tout en évitant d'introduire des biais provenant d'une simple imputation globale.

Algorithme:

```
Fonction replace_missing(dataset, attr):
    # Expression régulière pour les entiers
    integer_pattern = regex_compile(r'^[+-]?\d+$')
    # Expression régulière pour les décimaux
    decimal_pattern = regex_compile(r'^[+-]?\d*\.\d+$')

    # Remplacement des valeurs inappropriées par NaN
    Pour chaque valeur x dans dataset[attr]:
        Si x n'est pas NaN et ne correspond pas à integer_pattern ni decimal_pattern:
            x = NaN

    # Conversion de l'attribut en numérique (en supposant qu'il s'agit d'une colonne numérique)
    # Remplacement des NaN par la moyenne des instances appartenant à la même classe de fertilité
    Pour chaque classe de fertilité dans [0, 1, 2]:
        masque = dataset['Fertility'] égal à la classe de fertilité
        moyenne = moyenne des valeurs de dataset[attr] pour les instances satisfaisant le masque
        Remplacer les NaN dans dataset[attr] par la moyenne pour les instances satisfaisant le masque

    Retourner dataset
Fin de la fonction
```

Traitement des outliers

Dans le cadre de notre analyse des données sur la fertilité du sol, nous avons entrepris une étape cruciale de traitement des outliers. Les outliers, ou valeurs aberrantes, peuvent avoir un impact significatif sur nos résultats d'analyse, et il est essentiel de les traiter de manière appropriée. Nous avons exploré deux approches distinctes pour traiter ces valeurs aberrantes, à savoir le binning et le winsorization.
L'utilisation de ces deux méthodes a été entreprise dans le but de comparer l'impact des outliers sur la classification de nos données qui constitue la partie 2 de notre projet.

Méthode du Binning

La méthode du binning implique la division de l'ensemble de données en intervalles ou "bins". Cette approche offre une vue agrégée des données, réduisant ainsi l'impact des valeurs extrêmes.
Notre approche consiste à remplacer les valeurs situées en dehors des intervalles étudiés dans la partie **Analyse approfondie du jeu de données sur la fertilité du sol pour l'optimisation agricole** par la moyenne des valeurs de l'intervalle correspondant.
Voici comment nous avons mis en œuvre cette méthode :

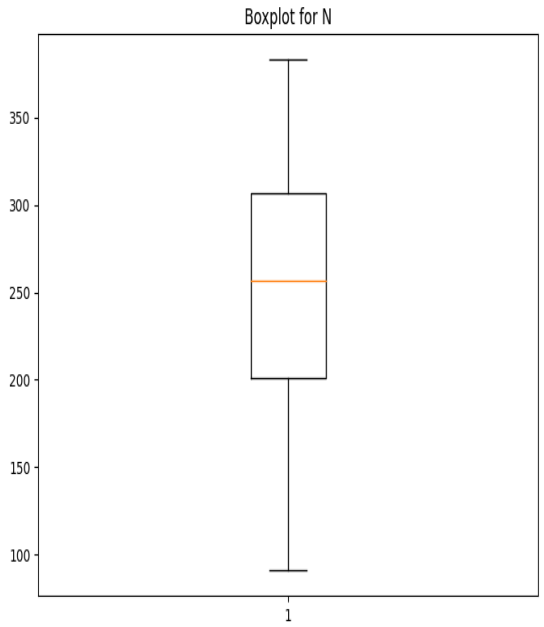
- Définition du nombre de bins (K) : Pour éviter de choisir arbitrairement le nombre de bins, nous avons utilisé la formule de Sturges , qui prend en compte la taille de l'échantillon, pour déterminer le nombre optimal de bins en fonction de la taille de notre ensemble de données.
- Catégorisation des données : Ensuite, nous avons découpé la variable cible en *k* intervalles. Chaque intervalle a été remplacé par la moyenne de ses valeurs.
- Application au Dataset : Les valeurs en dehors des intervalles ont été remplacées par la moyenne de l'intervalle correspondant.

Algorithme:

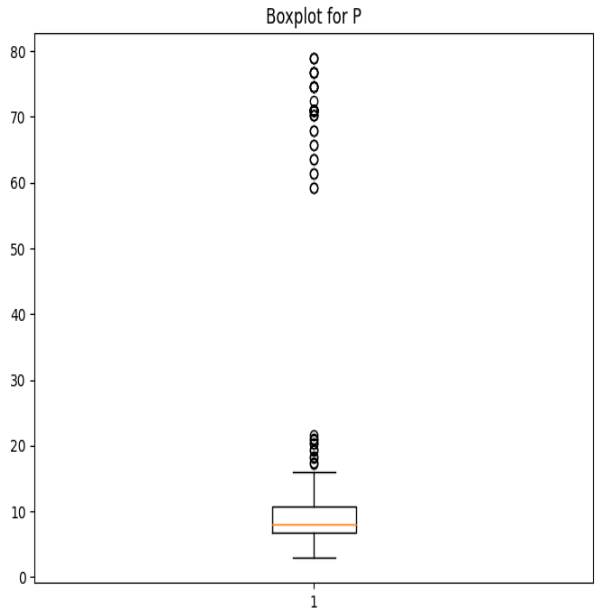
```
Fonction define_k(dataset):
    n = longueur(dataset)
    k = 1 + (10/3) * log10(n)
    Retourner plafond(k) - 1
Fin de la fonction

Fonction categorize(dataset, k):
    categories = liste_vide
    longueur_dataset = longueur(dataset)
    Pour i allant de 0 à k (non inclus):
        min_val_index = entier(longueur_dataset * i / k)
        max_val_index = entier(longueur_dataset * (i + 1) / k)
        intervalle = dataset[min_val_index:max_val_index]
        Ajouter intervalle à categories
    Retourner categories
Fin de la fonction

Fonction discretise(dataset, attr):
    dataset_trié = trier(dataset par attr)
    x = dataset_trié[attr]
    # Définir le nombre de plages K
    k = define_k(x)
    # Catégoriser
    categories = categorize(x, k)
    # Calculer la moyenne pour chaque catégorie et remplacer les valeurs par la moyenne
    Pour i, catégorie dans énumérer(categories):
        moyenne = catégorie.moyenne()
        categories[i] = Série_Pandas([moyenne] * longueur(catégorie), index=catégorie.index)
    # Combinez toutes les catégories dans une seule Série
    nouveau_dataset = concaténer(categories)
    dataset[attr] = nouveau_dataset
    Retourner dataset

Fonction binning(dataset, dataset2, min_threshold, max_threshold, attr):
    x = dataset[attr]
    masque = (x > max_threshold) | (x < min_threshold)
    dataset.loc[masque, attr] = dataset2.loc[masque, attr].astype(type(dataset[attr]))
    Retourner dataset
Fin de la fonction
```

Résultat du traitement

Nous remarquons que l'impact des outliers à été réduit sans pour autant s'en débarrasser complètement étant donné que ces valeurs aberrantes appartiennent à notre intervalle d'étude.



Graphe 1.16: boxplot pour N



Graphe 1.17: boxplot pour P

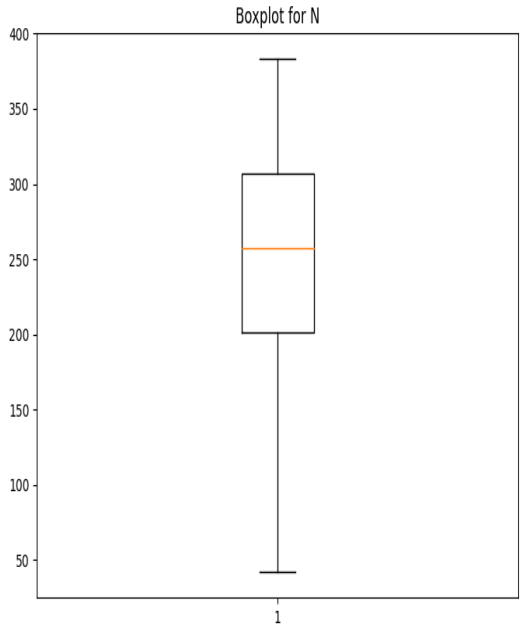
Méthode 2: La Winsorization

La deuxième méthode que nous avons utilisée est la winsorization. Cette technique consiste à remplacer les valeurs extrêmes d'une distribution par les valeurs limites de l'intervalle défini au travers du premier quartile et troisième quartile :

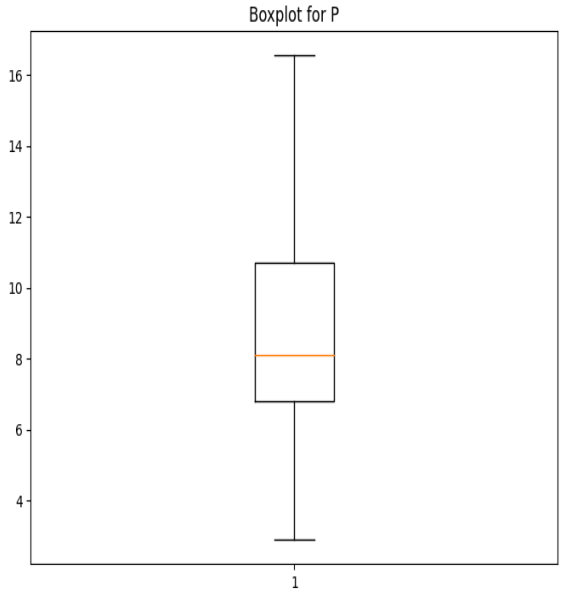
- Calcul des Valeurs Limites : Nous avons calculé les valeurs limites inférieure et supérieure en utilisant les percentiles.
- Remplacement des Outliers : Les valeurs aberrantes ont été remplacées par les valeurs limites correspondantes.

```
Fonction winsorize_outliers(dataset, attr):
    x = copie(dataset[attr])  # Faire une copie de la colonne
    q1 = x.quantile(0.25)
    q3 = x.quantile(0.75)
    val = 1.5 * (q3 - q1)
    below_q1 = q1 - val
    above_q3 = q3 + val

    # Remplacer les valeurs aberrantes
    x.loc[x < below_q1] = below_q1
    x.loc[x > above_q3] = above_q3

    dataset[attr] = x
    Retourner dataset
Fin de la fonction
```

Résultat du traitement
Nous remarquons que tous les outliers ont été remplacés par le minimum et maximum des intervalles du boxplot. Ceci pourrait avoir des effets indésirables sur notre classification.



Graphe 1.18: boxplot pour N



Graphe 1.19: boxplot pour P

Réduction des Données

Elimination des Redondances Verticales

Nous avons effectué une analyse pour identifier les lignes identiques dans le jeu de données. Le résultat montre qu'il y a 3 lignes identiques dans le jeu de données.
Nous avons procédé à la suppression des lignes identiques afin de réduire la redondance verticale dans le jeu de données. Les lignes supprimées sont les suivantes :

[377, 11.2, 636, 7.8, 0.6382022471910113, 0.54, 3.8, 0.37, 0.88, 0.31, 1.13, 0.85, 0.9288, 2]
[220, 8.6, 441, 7.43, 0.6382022471910113, 0.72, 11.7, 0.37, 0.66, 0.9, 2.19, 1.82, 1.2384, 0]
[270, 8.1, 636, 7.45, 0.55, 0.67, 10.2, 0.28, 0.44, 1.26, 7.75, 2.56, 1.1524, 1]

Elimination des Redondances Horizontales

L'analyse de corrélation a révélé des relations étroites entre certaines caractéristiques du sol. En particulier, les paires d'attributs suivantes présentent des corrélations significatives :

Matière Organique (OM) et Carbone Organique (OC) : Corrélation positive très forte de 1.
Azote (N) et Fertilité : Corrélation positive notable de 0.71.
Bore (B) et Manganèse (Mn) : Corrélation négative de -0.41.
Fer (Fe) et Bore (B) : Corrélation négative de -0.34.

Nous avons choisi de supprimer OM qui est parfaitement corrélées avec OC. Le choix de suppression d’autres attributs dépendra de l'étude que l’on mènera lors de la partie 2 du projet.

Normalisation des Données

Min-Max Normalisation

La méthode de normalisation Min-Max a été appliquée à chaque attribut du jeu de données. Cette méthode vise à ramener toutes les valeurs d'une caractéristique dans une plage spécifique, généralement entre 0 et 1. Les formules utilisées pour cette normalisation sont les suivantes :

$$normalized\_data = (x - min\_val) / (max\_val - min\_val) \times (new\_max - new\_min) + new\_min$$

où x est la valeur initiale, min_val et max_val sont respectivement les valeurs minimale et maximale de l'attribut, et new_min et new_max sont les bornes de la nouvelle plage.

Cette normalisation a pour effet de transformer toutes les valeurs d'une caractéristique dans une plage commune, ce qui facilite la comparaison entre différentes caractéristiques.

Z-Score Normalisation

La méthode de normalisation Z-Score, également appelée standardisation, a également été mise en œuvre pour chaque attribut du jeu de données. Cette méthode ajuste les valeurs d'une caractéristique en fonction de la moyenne et de l'écart-type de l'ensemble des données, selon la formule suivante :

$$normalized\_data = (data - mean) / stddev$$

Cette normalisation crée une distribution centrée autour de zéro, où la plupart des valeurs se situent dans une fourchette de ±3. Elle est utile lorsque les caractéristiques présentent des échelles différentes.

Ces deux méthodes de normalisation contribuent à rendre le jeu de données plus adapté aux algorithmes de machine learning, en atténuant les différences d'échelle entre les caractéristiques et en améliorant ainsi la performance des modèles.

Analyse et Prétraitement des Données Temporelles liées au COVID-19 aux États-Unis

Introduction

Le dataset 2 fournit une perspective temporelle détaillée sur la propagation du COVID-19 aux États-Unis, en se concentrant sur les cas, les tests et les taux de positivité par code postal (ZIPCODE) de 2019 à 2023. L'objectif est d'extraire des conclusions pertinentes à partir de ces données.

Exploration Initiale des Données

Avant d'entreprendre des analyses approfondies, commençons par explorer les caractéristiques principales du dataset.

Informations de Base sur le Dataset

Le dataset que nous avons examiné contient des informations liées à la propagation du COVID-19 aux États-Unis. Voici quelques informations de base sur ce dataset :

Aperçu des Caractéristiques du Dataset

Le dataset se compose de 337 entrées (lignes) et 11 colonnes (caractéristiques). Voici les caractéristiques présentes dans le dataset :

zcta (Zone Tabulation Code Area) : Code postal de la zone.
time_period : Période temporelle.
population : Population de la zone.
Start date : Date de début de la période.
end date : Date de fin de la période.
case count : Nombre de cas de COVID-19.
test count : Nombre de tests effectués.
positive tests : Nombre de tests positifs.
case rate : Taux de cas.
test rate : Taux de tests.
positivity rate : Taux de positivité.

Taille du Dataset

Le dataset occupe environ 29.1 KB en mémoire et a une forme de (337, 11), indiquant 337 lignes et 11 colonnes.

Types de Données

Le dataset comprend trois types de données principaux :

- float64 : Pour les valeurs numériques avec décimales.
- int64 : Pour les valeurs numériques entières.
- object : Pour les valeurs non numériques, généralement des dates.

Analyse des Données Manquantes

Certaines colonnes ont des valeurs manquantes, notamment "case count," "test count," et "positive tests." Ces valeurs manquantes devront être traitées lors du processus de nettoyage des données.

Nous avons effectué une analyse des valeurs manquantes pour chaque colonne du dataset. Voici un résumé des résultats :

Attr	zcta	time_period	population	start_date	end_date	case count	test count	positive tests	case rate	test rate	positivity rate
Nbr	0	0	0	0	0	26	12	27	0	0	0
%	0	0	0	0	0	7.72	3.56	8.01	0	0	0

Analyse des Caractéristiques des Attributs du Dataset de Fertilité du Sol

Analyse des tendances centrales

Nous avons calculé les mesures centrales suivantes pour chaque colonne du dataset temporel lié au COVID-19 :

Attr	zcta	time_period	population	case count	test count	positive tests	case rate	test rate	positivity rate
μ	94663.	43.69	50260.55	225.9	4938.12	380.2	19.39	454.84	5.83
Q2	950356	43	50477	104	4474	126	8.1	427	3
Mode	94085, 94086	21, 22,23,..., 67	23223, 50477	0	1295, 2251,2497,..., 6659	20, 47, 63	0	0.1	1.1
Symmetrie	N	N	N	N	N	N	N	N	N

Tableau 2.1: Tendances centrales du dataset 2

Analyse des quartiles

Nous avons calculé les quartiles pour chaque colonne du dataset temporel lié au COVID-19. Les quartiles fournissent des informations importantes sur la distribution des données et sont particulièrement utiles pour détecter la présence d'éventuels écarts importants dans les valeurs.

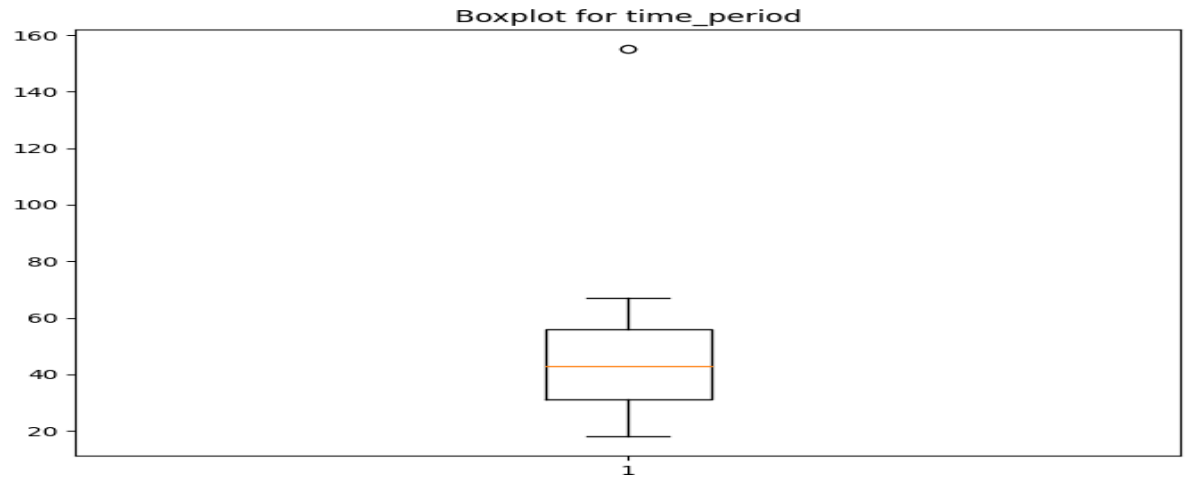
Attr	zcta	time_period	population	case count	test count	positive tests	case rate	test rate	positivity rate
Q0	94085	18	23223	0	11	11	0	0.1	0
Q1	94086	31	36975	42	2438	52	3.3	249.7	1.3
Q2	95035	43	50477	104	4474	126	8.1	427.1	3
Q3	95128	56	66256	316	6948	387	19.1	614.9	6.6
Q4	95129	155	79655	3627	20177	35000	260.7	1615.1	100

Tableau 2.2: Quartiles du dataset 2

Analyse des Boxplots et Valeurs Aberrantes

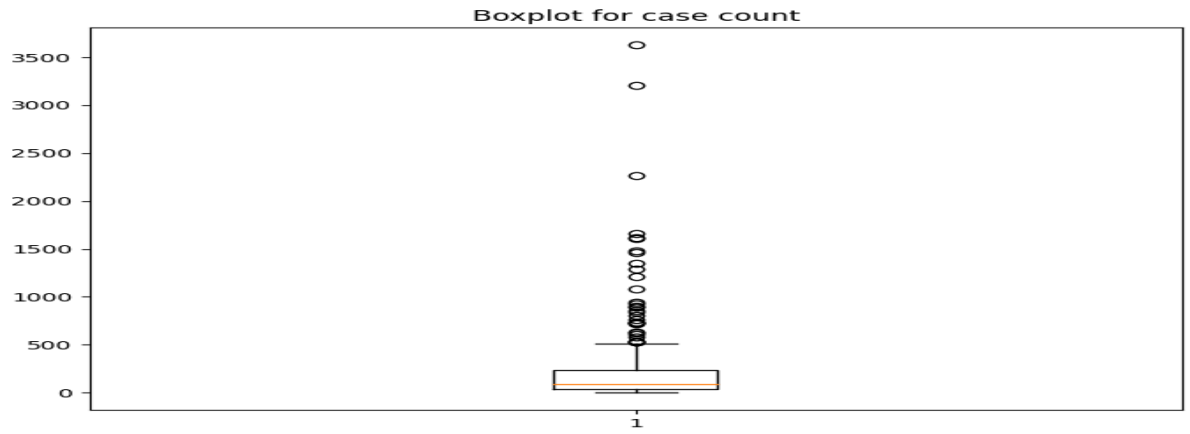
Nous avons créé des boxplots pour chaque attribut du dataset temporel lié au COVID-19. Les boxplots fournissent une vue visuelle des tendances centrales et de la dispersion des données. Les points aberrants détectés sont également notés.

Quelques Boxplots:



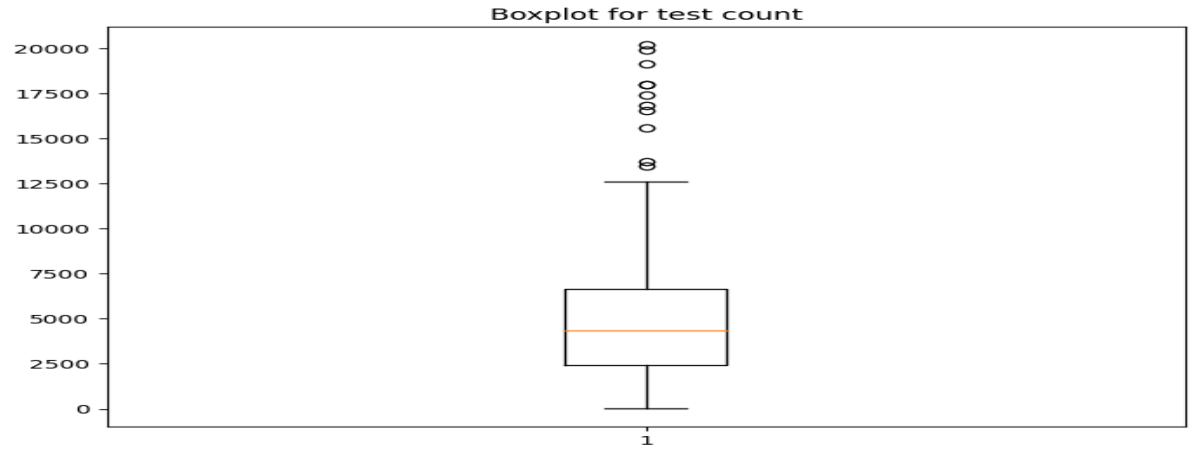
Graphe 2.1: boxplot pour time_period

time_period : Un point aberrant a été détecté à la valeur 155.



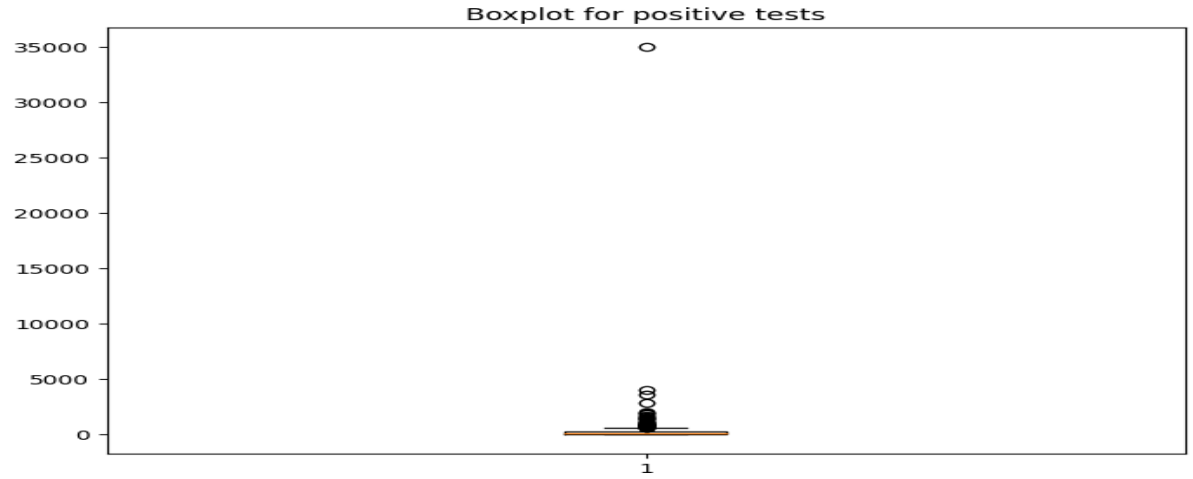
Graphe 2.2: boxplot pour case_count

case count : Plusieurs points aberrants ont été détectés avec des valeurs élevées, atteignant jusqu'à 3627.



Graphe 2.3: boxplot pour test_count

test count : Plusieurs points aberrants ont été détectés avec des valeurs élevées, atteignant jusqu'à 20177.



Graphe 2.4: boxplot pour positive_tests

positive tests : Plusieurs points aberrants ont été détectés avec des valeurs élevées, atteignant jusqu'à 35000.

Les colonnes case_rate, test_rate et positivity rate de notre dataset contiennent aussi plusieurs valeurs aberrantes.

Prétraitement du dataset

Le nettoyage des données est une étape cruciale pour garantir la qualité des analyses. Cela implique la gestion des valeurs manquantes, des données aberrantes et des incohérences.

Traitement des Valeurs Manquantes

Pour traiter les valeurs manquantes, nous avons remplacé chaque valeur manquante dans chaque colonne par la moyenne de cette colonne.Le code correspondant est le suivant :

```
FONCTION remplacer_valeurs_manquantes(ensemble_de_donnees, attribut)

    # Remplacer les valeurs NaN par la moyenne de la colonne

    valeur_moyenne = calculer_moyenne(ensemble_de_donnees[attribut].convertir_en_nombre())

    ensemble_de_donnees[attribut] =
ensemble_de_donnees[attribut].convertir_en_nombre().remplacer_nan_par(valeur_moyenne)

    # Vérifier les changements

    RETOURNER ensemble_de_donnees
```

Traitement des Données Aberrantes

Pour traiter les valeurs aberrantes, nous avons utilisé la méthode de winsorizing, qui remplace les valeurs aberrantes par les valeurs de seuil calculées à partir des quartiles. Le code correspondant est le suivant :

```
FONCTION traiter_outliers_boîte(ensemble_de_donnees, attribut)

    x = copie_de(ensemble_de_donnees[attribut])  # Faire une copie de la colonne

    q1 = calculer_quantile(x, 0.25)

    q3 = calculer_quantile(x, 0.75)

    valeur = 1.5 * (q3 - q1)

    en_dessous_q1 = q1 - valeur

    au_dessus_q3 = q3 + valeur

    # Remplacer les valeurs aberrantes par en_dessous_q1 ou au_dessus_q3

    x.loc[x < en_dessous_q1] = en_dessous_q1.convertir_en_type(x.type())

    x.loc[x > au_dessus_q3] = au_dessus_q3.convertir_en_type(x.type())

    ensemble_de_donnees[attribut] = x

    RETOURNER ensemble_de_donnees
```

Traitement des données Temporelle

Afin d'homogénéiser le format des dates dans les colonnes 'Start date' et 'end date' du dataset, nous avons créé une fonction convert_date qui utilise plusieurs formats de date possibles pour interpréter et reformater les dates. Cette approche est nécessaire car les dates peuvent être dans différents formats, par exemple, "Apr-07" ou "10/11/2020".

Cette fonction tente plusieurs formats de date jusqu'à ce qu'elle réussisse à interpréter la date correctement. Si le format de date inclut le mois en abrégé (par exemple, "Apr-07"), l'année 2023 est ajoutée, puis la date est reformulée dans le format souhaité "mm/dd/YYYY". Cette approche garantit que toutes les dates dans le dataset ont le même format.

Analyse et Prétraitement de Données Climatiques et Agricoles

Introduction

Le présent rapport vise à explorer les relations existantes entre les attributs climatiques (Température, Humidité, Précipitation), le type de sol, la végétation cultivée et l'utilisation d'engrais dans un dataset composé de 295 entrées.

Description du Dataset

Le dataset se compose de six colonnes, comprenant trois types d'attributs numériques et trois attributs catégoriels. Les attributs incluent la Température, l'Humidité et les Précipitations en tant que variables climatiques, le type de Sol, le type de Culture (Crop), et le type d'Engrais utilisé.

Prétraitement des Données

Le prétraitement des données est une étape cruciale pour garantir la qualité des résultats dans toute analyse de données. Nous avons adopté une approche de discrétisation pour rendre nos données plus accessibles et faciliter l'exploration des relations complexes entre les attributs.

Nous avons défini trois catégories de température : 'Low Temperature', 'Medium Temperature', et 'High Temperature'. Chaque observation a été assignée à l'une de ces catégories en fonction de sa position dans la distribution des températures dans le dataset.
De manière similaire, l'attribut d'humidité a été discrédité en 'Low Humidity', 'Moderate Humidity', et 'High Humidity'.
Les précipitations ont été divisées en 'Low Rainfall', 'Moderate Rainfall', et 'High Rainfall'.

Discrétisation en Classes d'Effectifs Égaux (Equal Frequency)

Nous avons choisi de discréditer notre dataset en utilisant la méthode des classes d'effectifs égaux pour les attributs de température, d'humidité et de précipitations. Cette approche vise à garantir une répartition équilibrée des données dans chaque catégorie. Les étapes suivantes ont été suivies pour chaque attribut :

	Temperature	Humidity	Rainfall	Soil	Crop	Fertilizer
0	Medium Temperature	Low Humidity	High Rainfall	Clayey	rice	DAP
1	High Temperature	High Humidity	Low Rainfall	laterite	Coconut	Good NPK
2	Low Temperature	Low Humidity	High Rainfall	silty clay	rice	MOP
3	Medium Temperature	High Humidity	Moderate Rainfall	sandy	Coconut	Urea
4	Medium Temperature	High Humidity	Low Rainfall	coastal	Coconut	Urea
...	...	...	...	...	...	...
290	Medium Temperature	High Humidity	Low Rainfall	sandy	Coconut	MOP
291	Medium Temperature	Moderate Humidity	Moderate Rainfall	silty clay	rice	MOP
292	Low Temperature	Moderate Humidity	High Rainfall	Clayey	rice	MOP
293	Medium Temperature	Moderate Humidity	Moderate Rainfall	Clayey	rice	MOP
294	Low Temperature	Moderate Humidity	High Rainfall	silty clay	rice	MOP

295 rows × 6 columns

Figure 2.1: Dataset discrétisé par la méthode d’effectifs égaux

Discrétisation en Classes d'Amplitudes Égales (Equal Width)

Cette approche de discrétisation divise l'ensemble de données en classes, chacune couvrant la même amplitude de valeurs. Cela permet une vision différente de la distribution des données.

	Temperature	Humidity	Rainfall	Soil	Crop	Fertilizer
0	Medium Temperature	Low Humidity	High Rainfall	Clayey	rice	DAP
1	High Temperature	High Humidity	Low Rainfall	laterite	Coconut	Good NPK
2	Low Temperature	Low Humidity	High Rainfall	silty clay	rice	MOP
3	Medium Temperature	High Humidity	Moderate Rainfall	sandy	Coconut	Urea
4	Medium Temperature	High Humidity	Low Rainfall	coastal	Coconut	Urea
...	...	...	...	...	...	...
290	Medium Temperature	High Humidity	Low Rainfall	sandy	Coconut	MOP
291	Medium Temperature	Low Humidity	Moderate Rainfall	silty clay	rice	MOP
292	Medium Temperature	Low Humidity	Moderate Rainfall	Clayey	rice	MOP
293	Medium Temperature	Low Humidity	Moderate Rainfall	Clayey	rice	MOP
294	Medium Temperature	Low Humidity	High Rainfall	silty clay	rice	MOP

295 rows × 6 columns

Figure 2.2: Dataset discrétisé par la méthode d’amplitudes égales

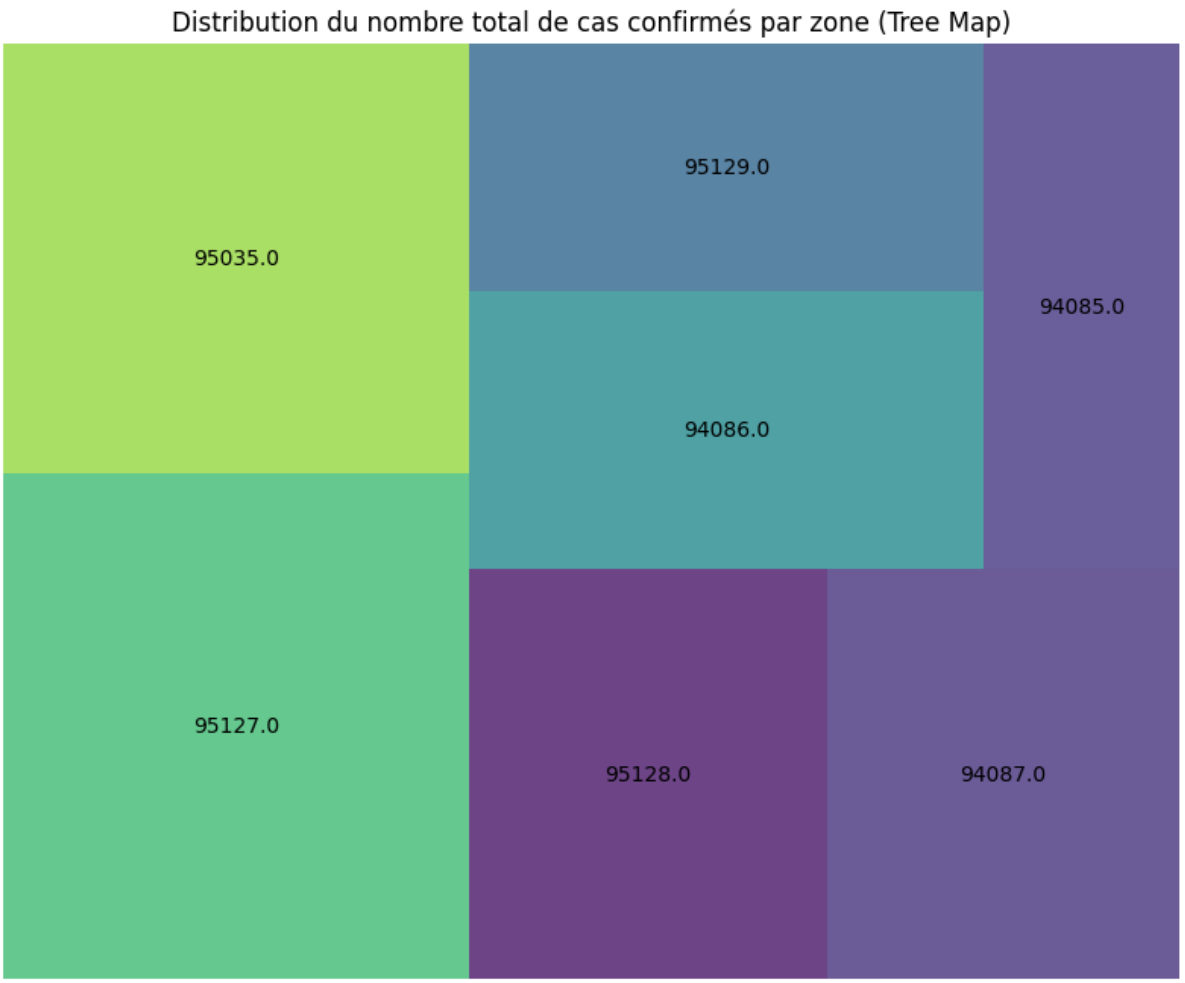
Partie II: Visualisations significatives des données temporelles

Introduction

Dans la continuité de notre exploration méthodique des données, la Partie II se focalise sur l'analyse temporelle à travers des visualisations significatives. Comprendre l'évolution des données au fil du temps est essentiel pour dégager des tendances, identifier des schémas récurrents et prendre des décisions éclairées.

Les graphiques utilisés dans cette section offrent une représentation visuelle de la distribution des cas confirmés et des tests positifs par zone, ainsi que l'évolution temporelle des tests COVID-19, des tests positifs, et du nombre de cas.

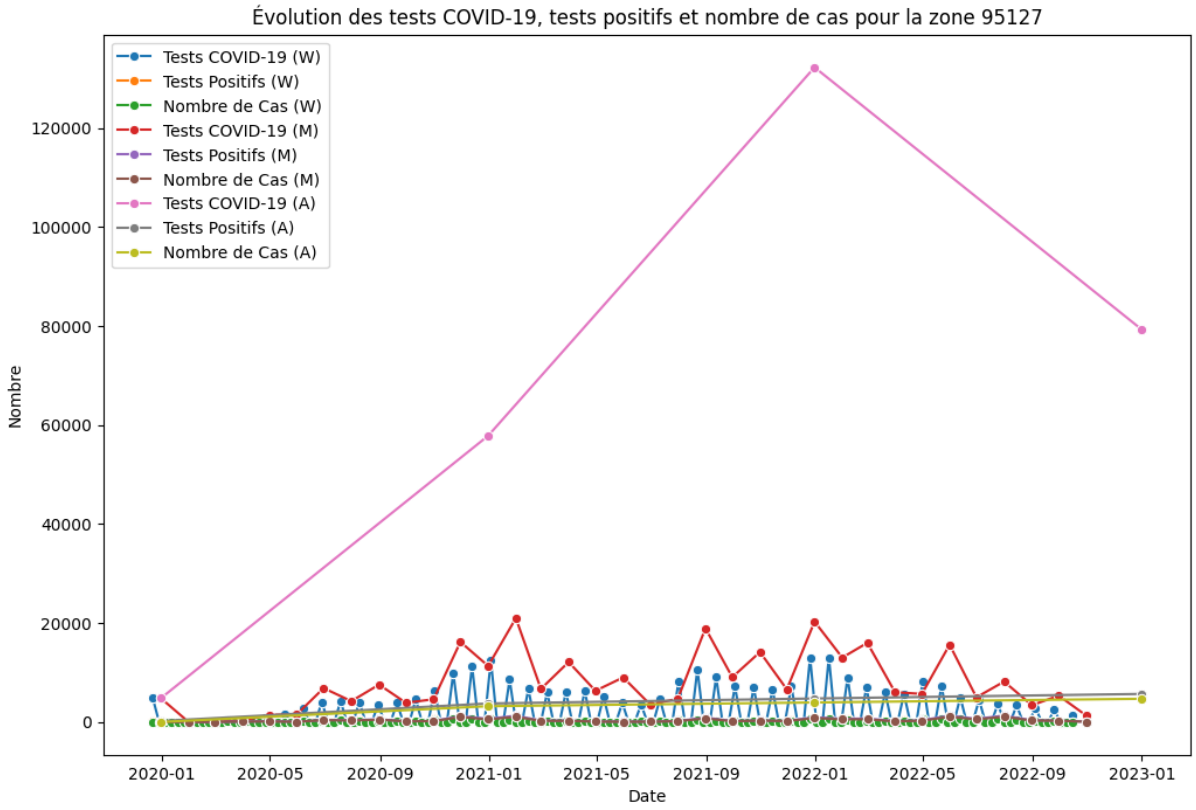
Requête 1 : Distribution du nombre total des cas confirmés et tests positifs par zones



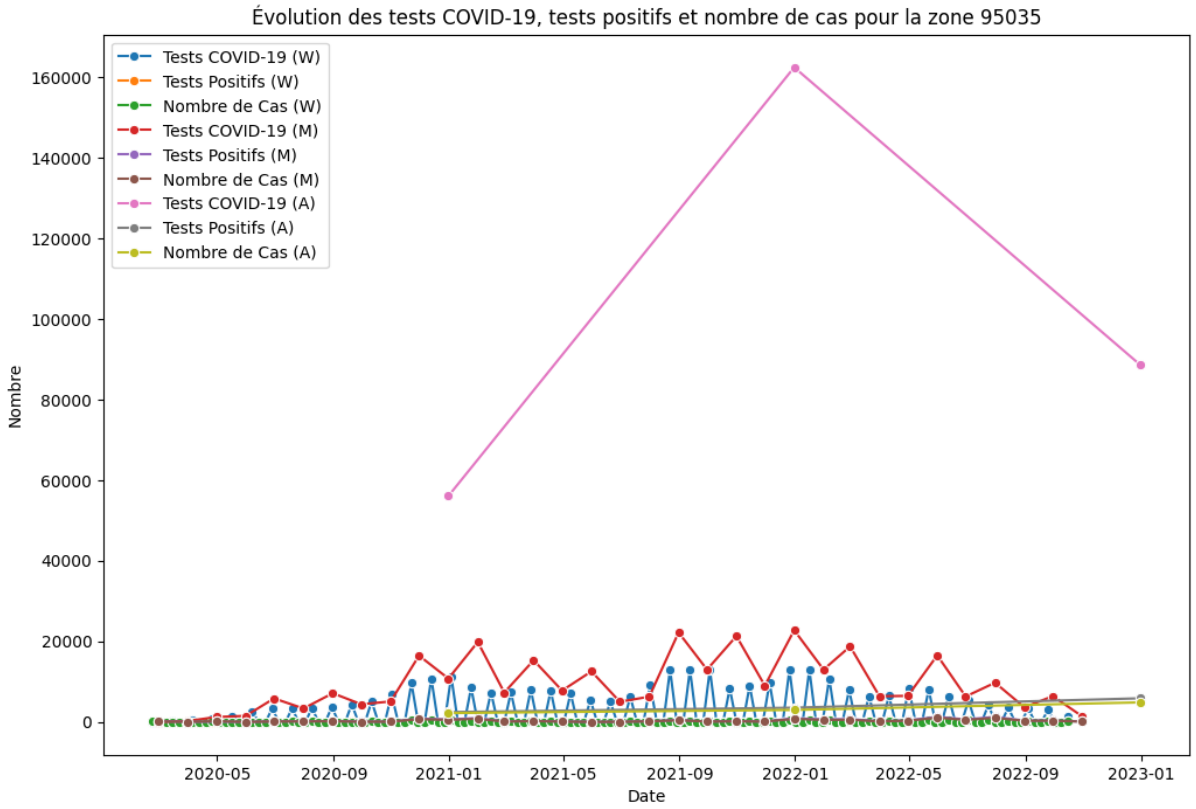
Distribution des Cas Confirmés et Tests Positifs par Zones (Tree Map/Bar chart)

Le graphique présente la répartition du nombre total de cas confirmés et de tests positifs par zone. Il est clair que la zone 95127 détient le plus grand nombre de cas confirmés de COVID-19, suivie de près par la zone 95035. Ces deux zones cumulent plus d'un tiers des cas confirmés dans l'ensemble des zones étudiées. Les autres zones présentent des chiffres presque équivalents, à l'exception de la zone 94085 qui affiche un nombre nettement inférieur.

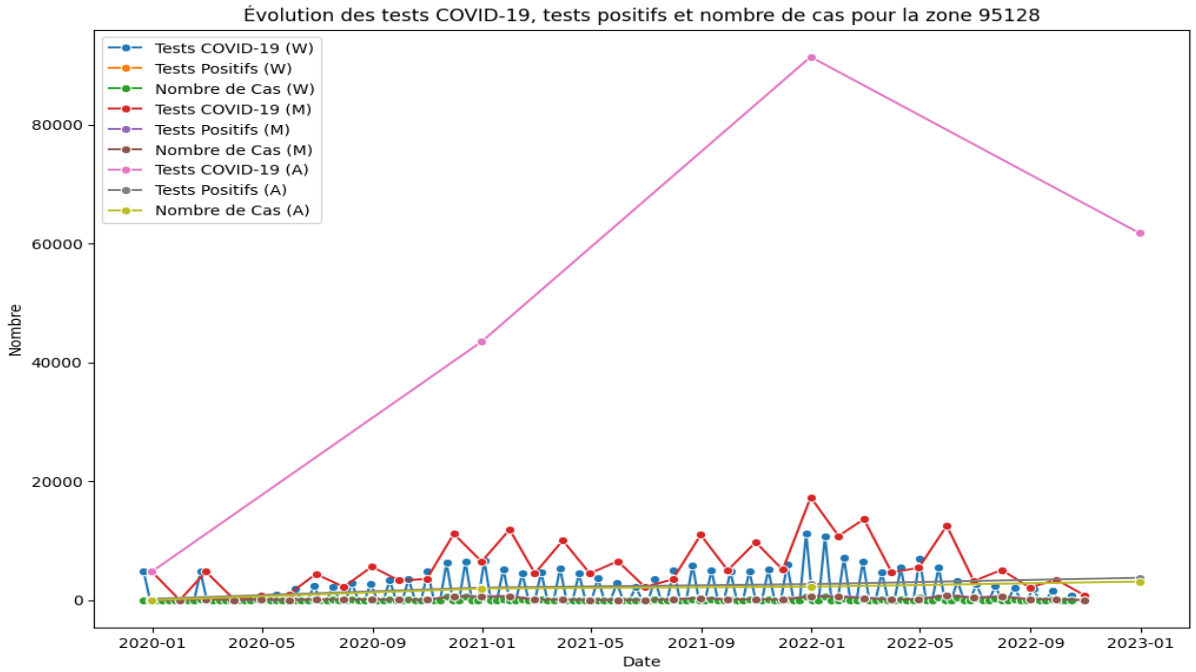
Requête 2 : Évolution des tests COVID-19, tests positifs et le nombre de cas évolué au fil du temps (hebdomadaire, mensuel et annuel) pour une zone choisie



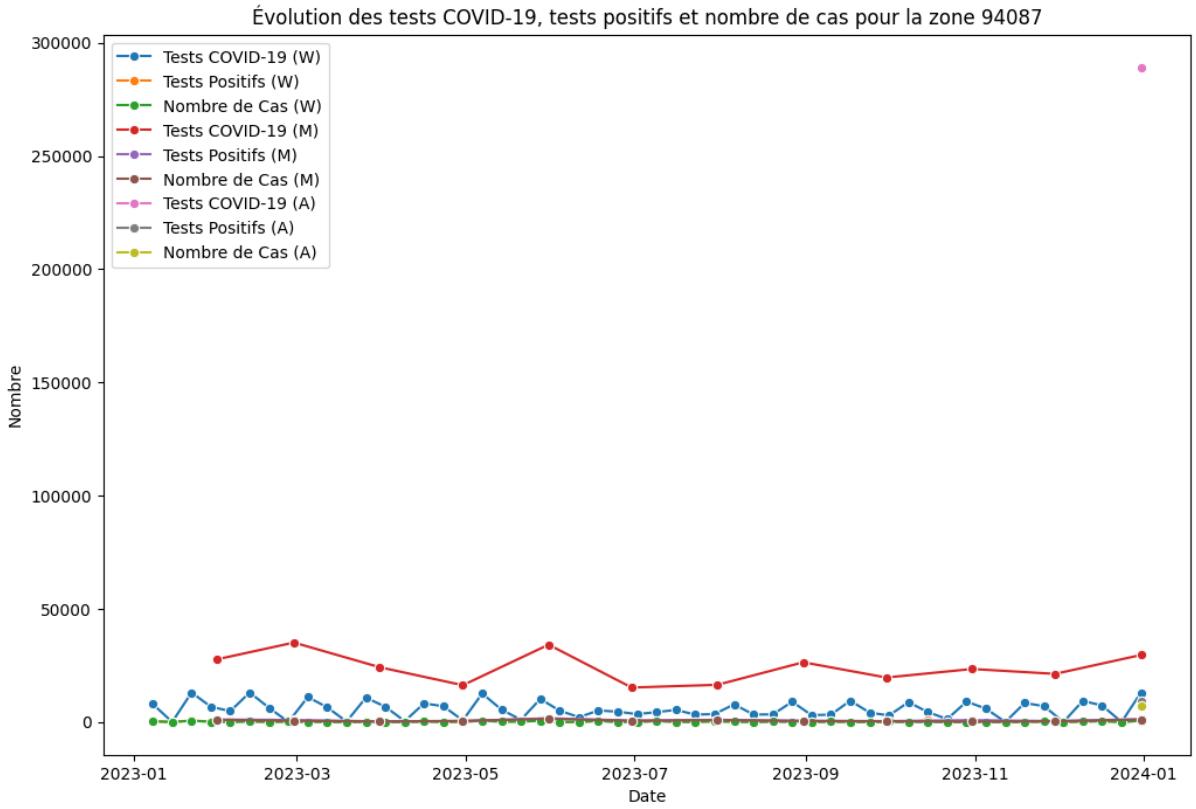
Graphe 2.5: Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 95127



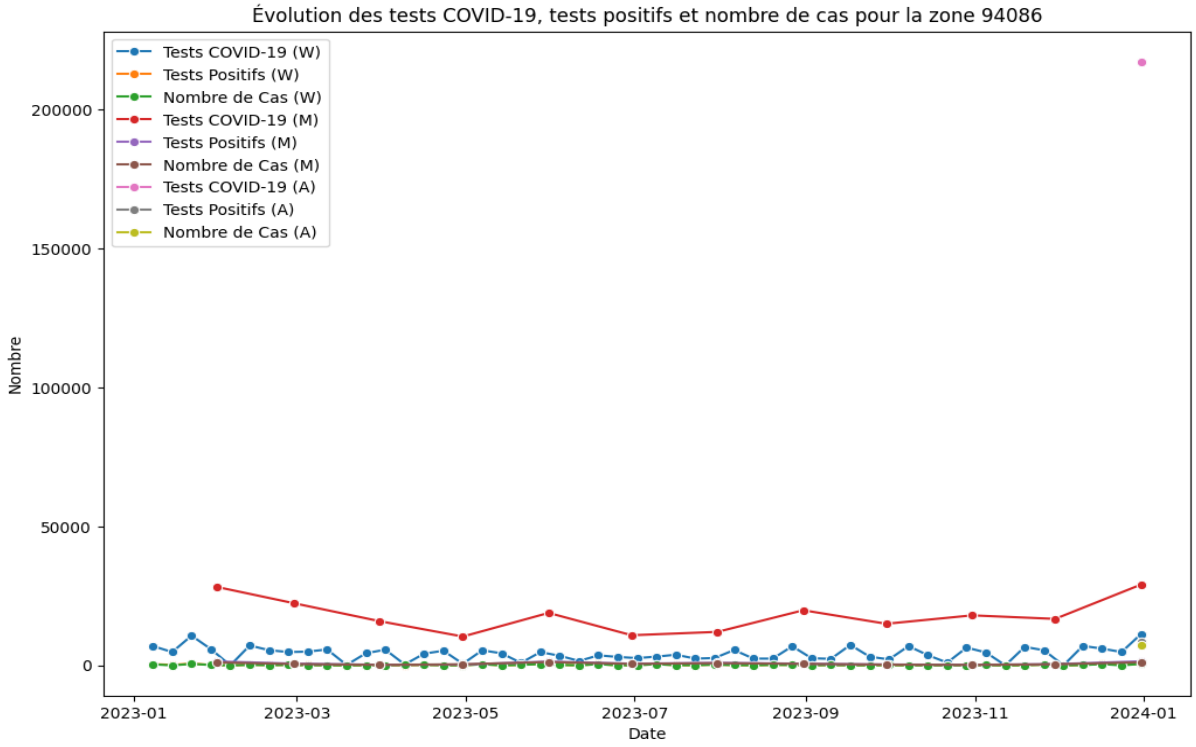
Graphe 2.6 : Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 95035



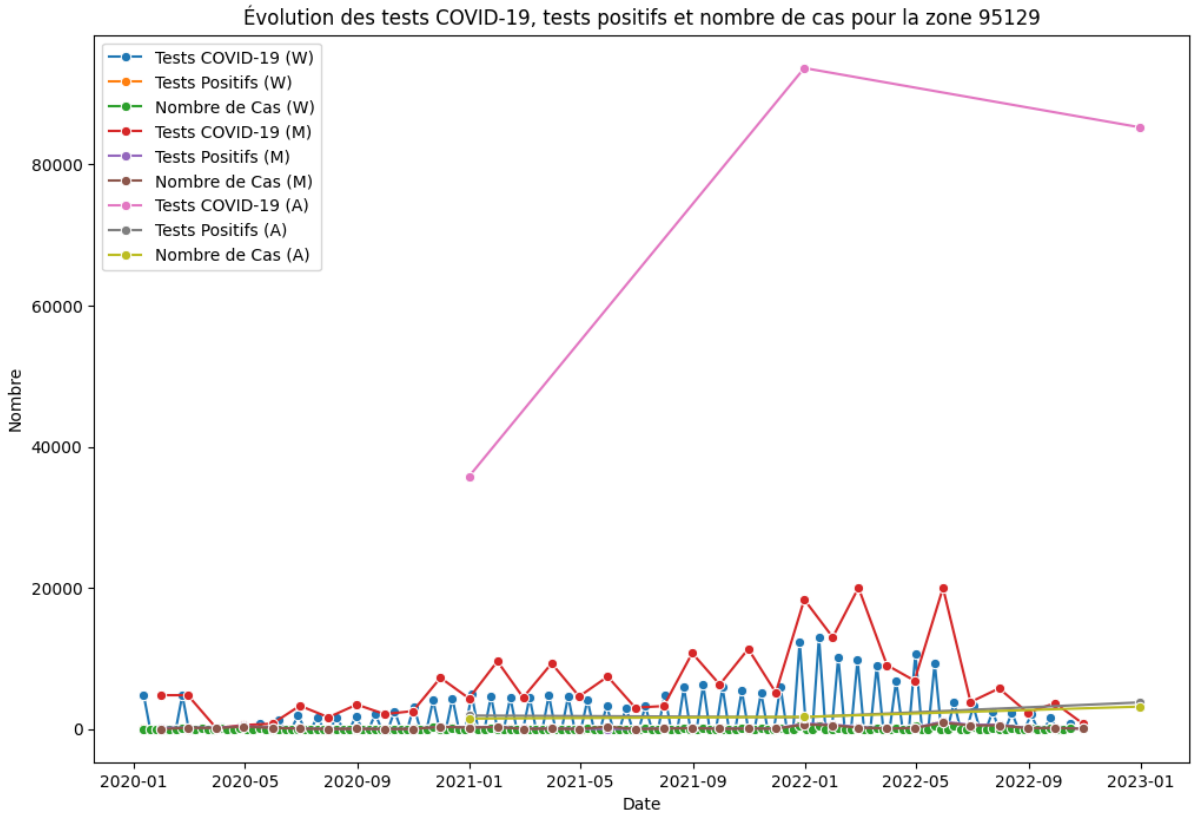
Graphe 2.7 :Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 95128



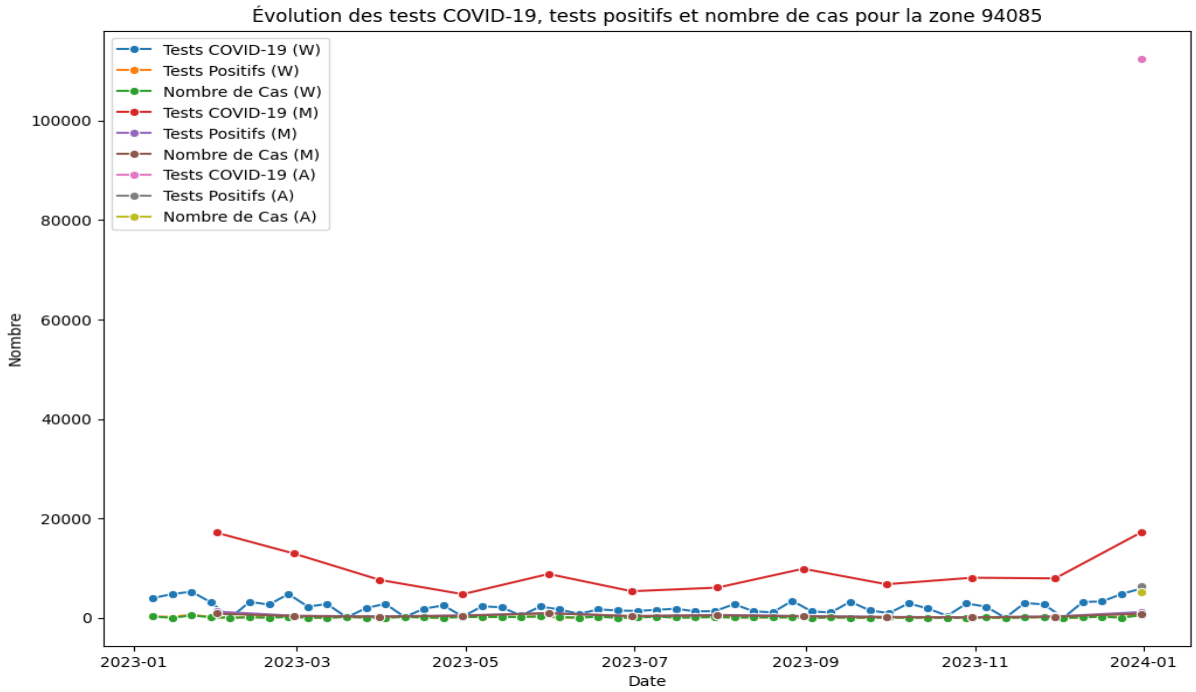
Graphe 2.8 :Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 94087



Graphe 2.9: Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 94086



Graphe 2.10 :Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 95129



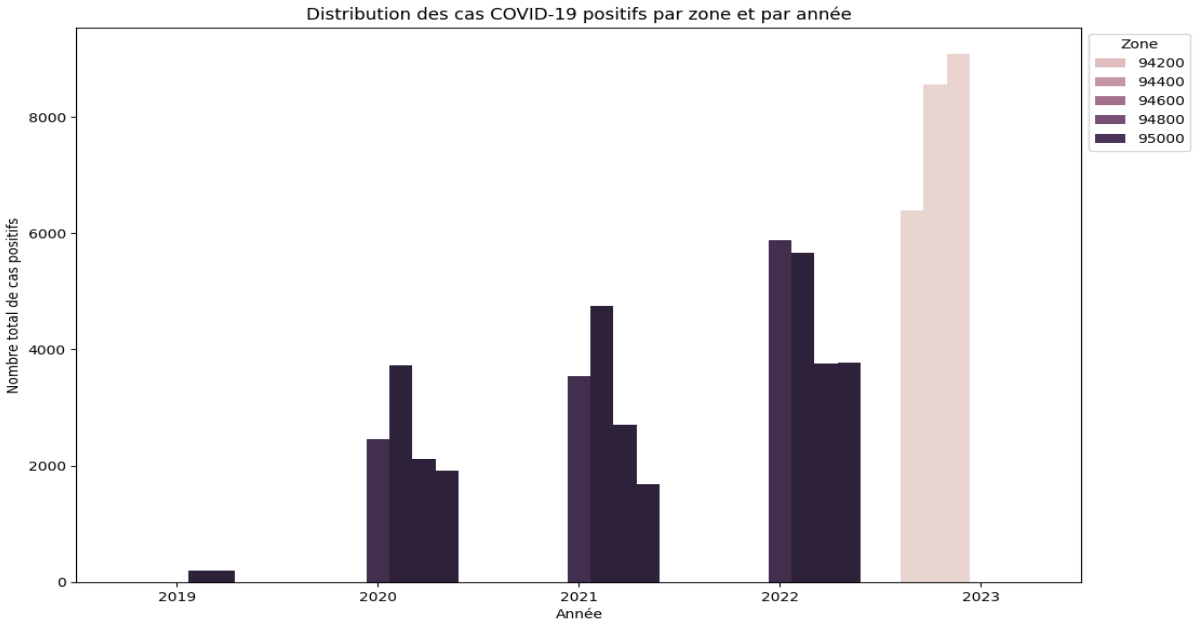
Graphe 2.11 :Évolution Temporelle des Tests COVID-19, Tests Positifs et Cas Confirmés pour la zone 94085

En analysant les tendances hebdomadaires, mensuelles, et annuelles, le graphique révèle une croissance significative du nombre de tests effectués dans la zone 95127 entre 2020 et 2023. Cette augmentation s'accompagne d'une croissance proportionnelle du nombre de cas et de tests positifs. Des schémas similaires sont observés dans les autres zones, soulignant que l'augmentation du nombre de tests révèle davantage de cas de COVID-19, bien que le nombre de cas et de tests positifs reste nettement inférieur au nombre total de tests effectués.

Par ailleurs, la répartition des données au fil du temps n'est pas uniforme entre les zones, comme illustré ci-dessous :

- Zone 95127 : Années 2019, 2020, 2021, 2022
- Zone 95035 : Années 2020, 2021, 2022
- Zone 95128 : Années 2019, 2020, 2021, 2022
- Zone 94087 : Année 2023
- Zone 94086 : Année 2023
- Zone 95129 : Années 2020, 2021, 2022
- Zone 94085 : Année 2023

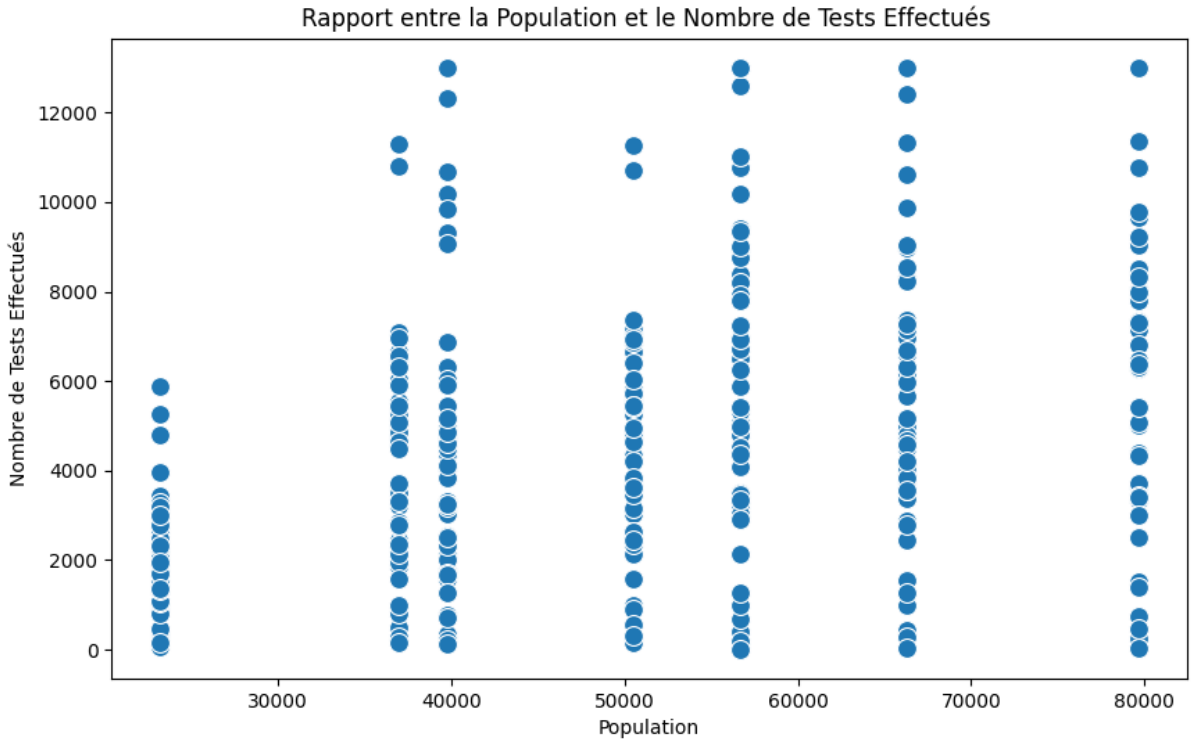
Requête 3 : Distribution des cas covid positifs par zone et par année



Graphe 2.12 :Répartition des Cas COVID-19 Positifs par Zone et par Année (Stacked Bar chart)

Le graphe ci-dessus offre un aperçu détaillé de la répartition des cas COVID-19 positifs par zone et par année. Il est à noter l'absence de données pour certaines années, notamment pour les zones 94085, 94086, et 94087. Une augmentation du nombre de cas positifs est également observée d'une année à l'autre, avec l'année 2023 affichant le plus grand nombre de cas, principalement concentré dans les zones 94085, 94086, et 94087.

Requête 4 : Graphique du rapport entre la population et le nombre de tests effectués



Graphe 2.13 : Rapport entre la Population et le Nombre de tests effectués

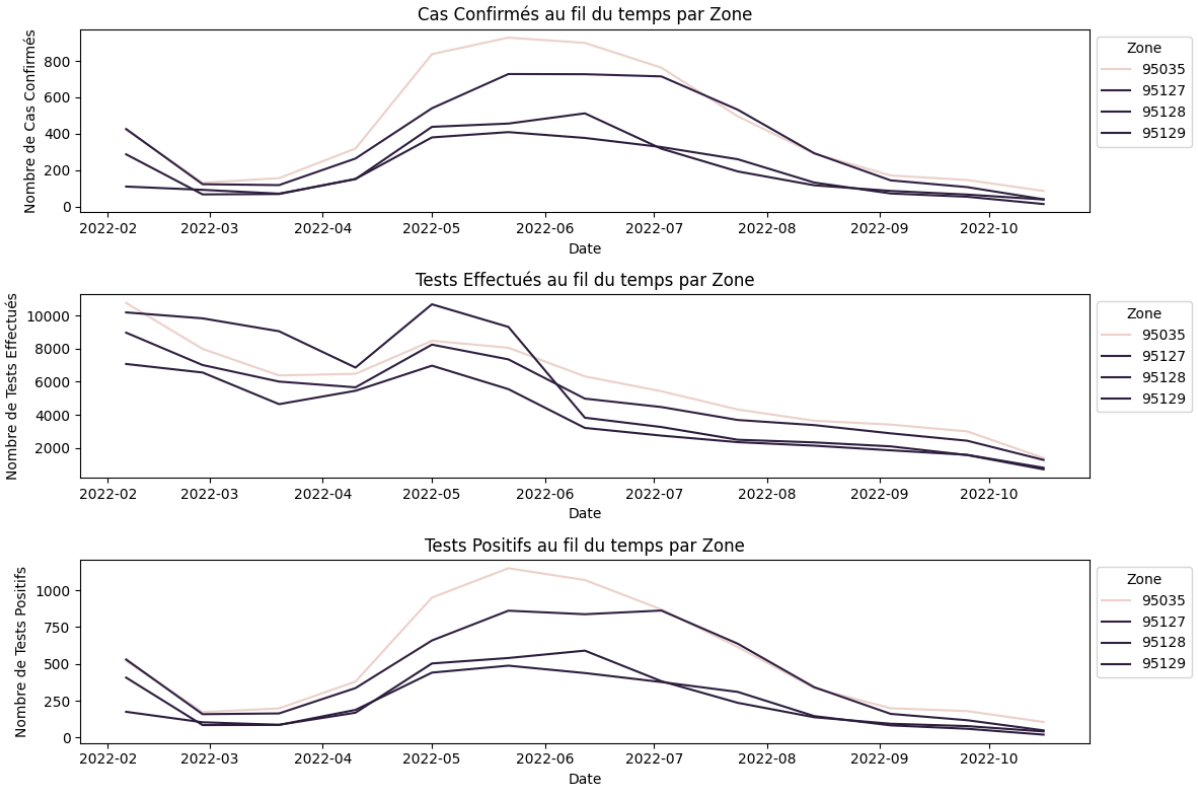
Le scatter plot est le meilleur choix graphique pour représenter efficacement la relation entre la population et le nombre de tests car il permet une visualisation détaillée des données individuelles, identifie des schémas, clusters et tendances, détecte des relations non linéaires, et offre une interprétation intuitive. Sa capacité à représenter visuellement chaque point par rapport aux axes de la population et du nombre de tests en fait un outil idéal pour une analyse approfondie et accessible à un large public.

Requête 5 : Les cinqs zones les plus fortement impactées par le coronavirus

Les cinq zones les plus impactées par le coronavirus sont représentés par un tableau :

zcta	case count
95127	19816.0
95035	15335.0
95128	8778.0
94087	8124.0
94086	7749.0

Requête 6 : Rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour chaque zone (pour une période de temps choisie)



Graphe 2.14 : Rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour chaque zone (pour une période de temps choisie)

Nous observons que, pour cette période, les cas confirmés et les tests positifs affichent une corrélation positive, montrant une proportionnalité au fil du temps. La courbe des tests effectués suit une tendance similaire, bien que sa diminution soit plus prononcée vers juin 2022.

Plusieurs hypothèses peuvent être formulées à partir de ces observations. Il semble y avoir une croissance significative des cas confirmés, suivie d'une augmentation des tests effectués, dont beaucoup se sont révélés positifs. Toutefois, avec le temps, le nombre de tests a diminué, peut-être pour des raisons économiques ou en raison d'une insuffisance de tests disponibles. Malgré cette diminution, les cas confirmés et les tests positifs restent importants. Entre août 2022 et octobre 2022, les graphiques indiquent une décroissance.

Il est important de souligner que ces hypothèses doivent être formulées par des experts du domaine ayant une connaissance approfondie du contexte pour assurer une interprétation précise des tendances observées.

Partie III : Système de recommandation à l'aide de l’algorithme Apriori

Introduction

La Partie III de notre étude se consacre à l'exploration et à la mise en œuvre d'un système de recommandation innovant, utilisant l'algorithme Apriori. Cette approche révolutionnaire repose sur des principes de fouille de données fréquentes, permettant ainsi de découvrir des motifs d'association entre différentes entités au sein d'un ensemble de données.

Les systèmes de recommandation jouent un rôle essentiel dans le domaine de l'information personnalisée, en fournissant des suggestions pertinentes aux utilisateurs en fonction de leurs comportements passés, de leurs préférences et de leurs interactions. L'algorithme Apriori, initialement conçu pour l'analyse de paniers de marché, s'est avéré particulièrement efficace dans le contexte des systèmes de recommandation.

Au cours de cette partie, nous explorerons en détail le fonctionnement de l'algorithme Apriori et son adaptation à la construction d'un système de recommandation. Nous aborderons la génération de règles d'association, la mesure de la confiance et du support, ainsi que l'application pratique de ces règles pour formuler des recommandations personnalisées. Cette approche permettra de créer un système de recommandation intelligent capable d'anticiper les besoins et les préférences des utilisateurs, améliorant ainsi leur expérience et la pertinence des suggestions fournies.

Création des transactions

L'identification des transactions et des itemsets dans un dataset implique une bonne compréhension de la structure des données et de les représenter en conséquence.
Quelques questions à se poser peuvent être :
Quelles combinaisons d'articles ont-elles une signification spécifique ?
Que veut-on recommander ? Certains ensembles d'articles ont-ils une signification spécifique ?
Dans le contexte de notre dataset, des éléments immuables tels que les conditions météorologiques (température, humidité, précipitations) et la qualité du sol peuvent être considérés comme des transactions. Parallèlement, la culture et les types d'engrais peuvent être considérés comme des ensembles d'articles.

Notre objectif consiste à déterminer la meilleure recommandation de culture et d'engrais pour une nouvelle entrée d'information.Cette association peut être utile aux agriculteurs qui souhaitent savoir quoi planter dans leurs champs et avec quel engrais.

L'algorithme Apriori

L'algorithme APriori est un algorithme d'exploration de données conçu par Rakesh Agrawal et Ramakrishnan Sikrant, dans le domaine de l'apprentissage des règles d'association. Il sert à retirer des ensembles de tous les itemsets qui reviennent fréquemment dans un dataset [15].

```
Pseudo-Algorithme
Entrée:k:entier, suppmin:entier
Sortie:L : liste
Variables:
        Ck:liste des candidats de l'itemset (K)
        Lk:liste des itemsets frequents(k)
Début
L={itemsets frequents de taille 1}
Tant que ( Lk!=0 ):
        Ck+1=generate_candidats(Lk)
        Pour t dans Ck+1 :
                Calcul support(t)
                Si support(t)>suppmin
                        Ajouter (t,Lk+1)
         L= L U Lk+1

Retourner LULK+1

Fin
```

Cet algorithme effectue d'abord une transformation du dataset d'origine en un format approprié, appelé format transactionnel. Cela implique de convertir le dataset en une structure de données composée de deux colonnes : une colonne pour les transactions et une colonne pour les itemsets. Dans notre cas, la première colonnes correspond aux conditions : Température Humidity,Rainfall,Soil. La deuxième colonne items est constituée de Crop et Fertilizer.

Ensuite, pour chaque candidat présent dans l'ensemble de transactions. Calculer son support et l'ajouter à L1 si le support est supérieur ou égale au support minimum (minsup).

Répéter la génération des ensembles candidats de taille k en combinant les ensembles fréquents existants, et le filtrage les ensembles candidats en ne conservant que ceux dont tous les sous-ensembles sont fréquents.
Continuer le processus de génération d'ensembles fréquents jusqu'à ce qu'aucun nouvel ensemble fréquent ne puisse être généré.

Les ensembles de fréquents obtenus représentent les articles qui co-occurrent fréquemment dans les transactions

Génération des règles d’association

À partir des ensembles d’itemsets fréquents obtenus, générez toutes les règles possibles en divisant chaque ensemble en deux parties (antécédent et conséquent).

Un ensemble de taille k génère 2^k règles d’associations

Les règles générées doivent satisfaire une certaine mesure de confiance minimale.

Les règles d'association obtenues décrivent les relations entre les articles, indiquant à quel point la présence d'un article (antécédent) est liée à la présence d'un autre article (conséquent).

Explication des mesures de corrélations

L'analyse des règles d'association est un aspect crucial de l'exploration de données transactionnelles, visant à découvrir des relations intéressantes entre les articles. Pour évaluer la qualité et la force de ces règles, différentes mesures de corrélation sont utilisées. Dans ce rapport, nous présentons plusieurs de ces mesures et discutons de leurs avantages et limites.

Mesures de Corrélation

1. Confidence

La Confidence offre une vue directe de la probabilité conditionnelle, facilitant la compréhension de la force de la relation entre l'antécédent et le conséquent. C'est une mesure intuitive et simple à interpréter. Cependant,elle ne tient pas compte de la fréquence globale des articles dans le dataset, ce qui peut entraîner des évaluations biaisées si certains articles sont très fréquents.

```
Algorithme calculate_confidance
Entrée : L:liste
Sortie :regle :regles d’association
Variables: antecedent,consequent
Début

        Antecedent,consequent = regle
        support_antecedent=calculsupport(antecedent)
        support_regle=calculsupport(regle)
        confiance =support_regle/support_antecedent

        Retourner confiance

Fin
```

2. All-Confidence

L'All-Confidence normalise la confidence par rapport au maximum des supports individuels des antécédents et des conséquents. Cela permet de prendre en compte la distribution globale des articles dans le dataset.
Néanmoins, elle peut ne pas refléter la véritable importance d'une règle dans des cas où un seul item est très fréquent.

```
Algorithme calculate_all_confidence
Entrée : L:liste
Sortie :regle :regles d'association
Variables: antecedent,consequent
Début

        Antecedent,consequent = regle
        support_antecedent=calculsupport(antecedent)
        support_consequent=calculsupport(consequent)
        support_regle=calculsupport(regle)

        max_support= max (support_antecedent, support_consequent)
        all_confiance =support_regle/max_support

        Retourner all_confiance

Fin
```

3. Max-Confidence

Max-Confidence prend en compte le maximum entre la confiance conditionnelle et la confiance inversée, offrant une vision bidirectionnelle de la force de la règle. Cela permet d'identifier des relations fortes, peu importe la direction antécédent → conséquent ou conséquent → antécédent.
Cependant, dans certains cas, cette approche peut ne pas rendre compte de manière équilibrée des deux directions de la règle.

```
Algorithme calculate_max_confidence
Entrée : L:liste
Sortie :regle :regles d'association
Variables: antecedent,consequent
Début

        Antecedent,consequent = regle
        support_antecedent=calculsupport(antecedent)
        support_consequent=calculsupport(consequent)
        support_regle=calculsupport(regle)

        max_confidance= max (support_regle/support_antecedent,
support_regle/support_consequent)


        Retourner max_confidance

Fin
```

4. Cosine Similarity

La Cosine Similarity mesure la similarité entre les ensembles d'articles comme des vecteurs, ce qui peut être avantageux lorsque la magnitude des ensembles est importante. Elle est également adaptée à des ensembles de tailles variables.
Elle peut être sensible aux différences de magnitude entre les ensembles, et ne prend pas en compte le chevauchement exact des articles.

```
Algorithme calculate_cosine_similarity
Entrée : L:liste
Sortie :regle :regles d'association
Variables: antecedent,consequent
Début

        Antecedent,consequent = regle
        support_antecedent=calculsupport(antecedent)
        support_consequent=calculsupport(consequent)
        support_regle=calculsupport(regle)

        cosine_similarity= support_regle/racine (support_antecedent*
support_consequent)


        Retourner cosine_similarity

Fin
```

5. Jaccard Similarity

La Jaccard Similarity mesure la similarité entre les ensembles d'articles en termes de chevauchement. Elle est calculée en divisant le support combiné par la taille de l'union des supports des antécédents et des conséquents.
Cette mesure fournit une mesure simple et intuitive de la similarité en termes de chevauchement d'articles, ce qui la rend facile à comprendre et à interpréter.
Cependant, elle peut ne pas être sensible aux variations dans la fréquence des articles.

```
Algorithme calculate_jaccard_similarity
Entrée : L:liste
Sortie :regle :regles d'association
Variables: antecedent,consequent
Début

        Antecedent,consequent = regle
        support_antecedent=calculsupport(antecedent)
        support_consequent=calculsupport(consequent)
        support_regle=calculsupport(regle)

        jaccard_similarity= support_regle/(support_antecedent +
support_consequent -  support_regle )


        Retourner jaccard_similarity

Fin
```

6. Kulczynski Similarity

La Kulczynski Similarity, basée sur la moyenne des confiances conditionnelles, offre une mesure équilibrée de la force de la règle en considérant à la fois l'antécédent et le conséquent. Elle peut être plus robuste lorsque la confiance conditionnelle seule est biaisée. Cependant, elle peut être influencée par des variations dans la taille des ensembles et peut nécessiter une attention particulière dans l'interprétation.

```
Algorithme calculate_kulczynski_similarity
Entrée : L:liste
Sortie :regle :regles d'association
Variables: antecedent,consequent
Début

        Antecedent,consequent = regle
        support_antecedent=calculsupport(antecedent)
        support_consequent=calculsupport(consequent)
        support_regle=calculsupport(regle)

        kulczynski_similarity= 0.5*(support_regle/support_antecedent +
support_regle/support_consequent)


        Retourner kulczynski_similarity

Fin
```

Chaque mesure a ses avantages et ses inconvénients, et le choix dépend des objectifs spécifiques de l'analyse des règles d'association et des caractéristiques des données sous-jacentes

7-Mesure du lift

La mesure du Lift évalue la force d'association entre l'antécédent et le conséquent, offrant des informations cruciales sur la significativité d'une règle d'association. Un Lift supérieur à 1 indique une association positive, ce qui est utile pour identifier des relations significatives.

Cependant, puisque le Lift est calculé a partir de la confiance, lui aussi peut être sensible à la fréquence absolue des articles.

```
Algorithme calculate_lift_measure
Entrée : L:liste
Sortie :regle :regles d'association
Variables: antecedent,consequent
Début

    Antecedent,consequent = regle
    support_antecedent=calculsupport(antecedent)
    support_consequent=calculsupport(consequent)
    support_regle=calculsupport(regle)

    confidence=calculate_confidence(regle,L)
    Lift_measure = confidence / support_consequent

    Retourner lift_measure

Fin
```

Chaque mesure présente des avantages et des limites. La Confidence offre une interprétation directe, tandis que l'All-Confidence et la Max-Confidence fournissent des perspectives normalisées et bidirectionnelles. La Cosine et la Jaccard Similarity utilisent des approches différentes pour mesurer la similarité entre ensembles d'articles. La Kulczynski Similarity offre une moyenne équilibrée des confiances conditionnelles.

Cependant, il est important de noter que ces mesures peuvent ne pas tenir compte de la taille des ensembles fréquents, et certaines peuvent ne pas être sensibles aux valeurs nulles dans les données.

En conclusion, le choix de la mesure de corrélation dépend du contexte spécifique de l'application et des objectifs d'analyse. Chacune de ces mesures offre une perspective unique sur la force des règles d'association extraites de données transactionnelles.

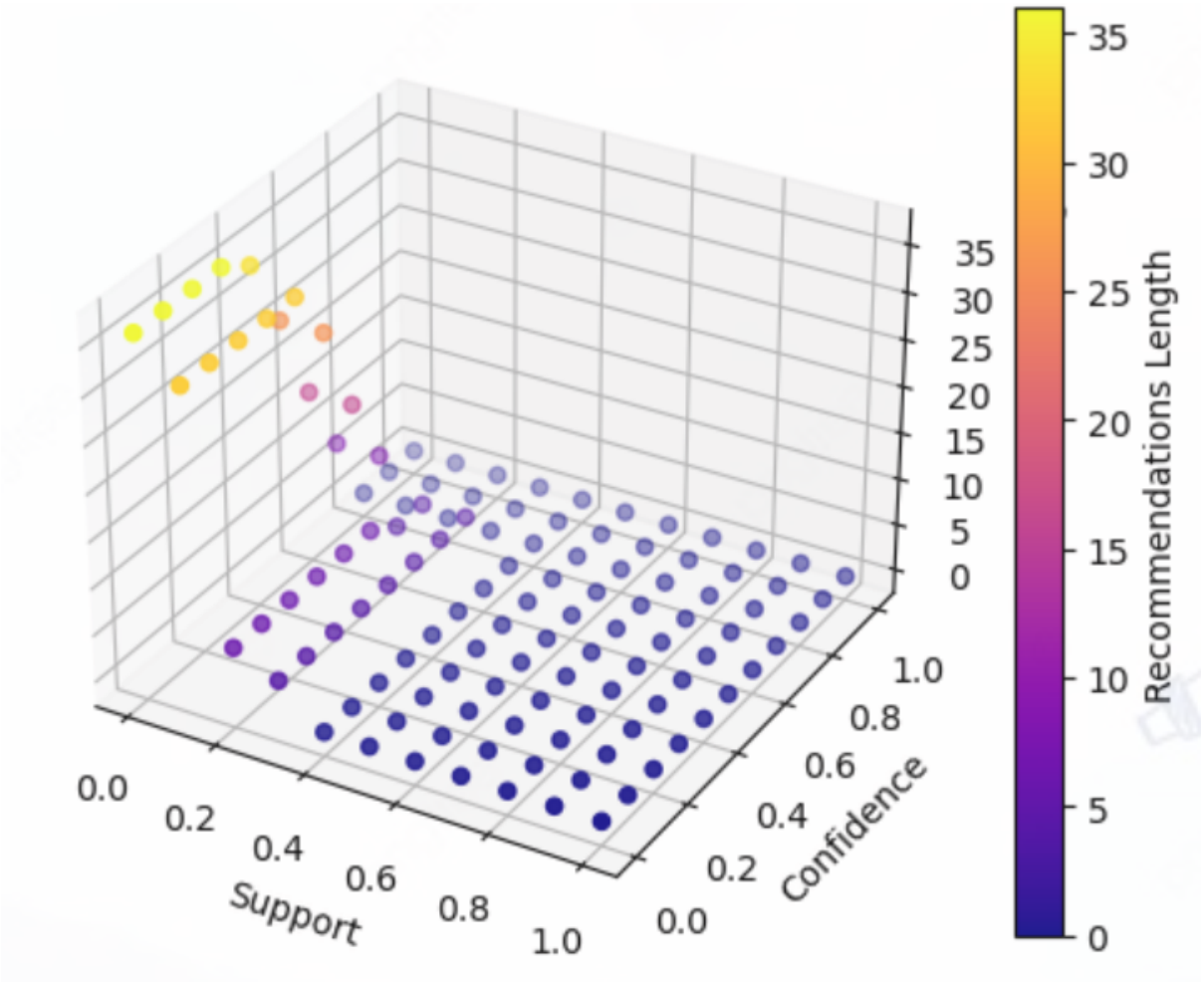
Extraction des fortes règles d’association

Pour l’extraction des fortes règles d’association, il suffit de générer toutes les règles d’association, puis les filtrer en ne gardant que les règles avec une confiance supérieure à confmin .

```
Algorithme fortes_regles_association
Entrée :
Min_confidence: entier ; L : liste ; associationn_rules :regles
d'association
Sortie :regle :regles d'association
Variables: fortes_regles : liste
Début

        Pour regle dans associationn_rules:
                Confidence = mesure_correlation(regle,L)
                Si Confidence >= Min_confidence
                Alors
                        fortes_regles.append(regle)
                fsi;
        Finpour;

        Retourner fortes_regles

Fin
```

Résultat des variations de minsup et minconf

Dans cette section, nous présentons les résultats de notre analyse en variant les valeurs de support minimum (minsup) et de confiance minimum (minconf) dans le cadre de l'algorithme d'extraction de règles d'association, en l'occurrence l'algorithme Apriori. L'objectif est d'observer comment ces variations influent sur le nombre de règles d'association fortes générées.
Dans notre analyse, nous avons observé une relation inverse entre le nombre de règles d'association fortes générées et les valeurs de minsup et minconf. Cette tendance est logique, car des valeurs plus élevées de minsup entraînent la sélection de moins de candidats en raison de leur faible support, et le même principe s'applique à la confiance.



Graphe 3.1 : Scatter plot de la taille de la recommandation par rapport a minsup et minconf

Il est intéressant de noter que lorsque minsup > 0.5 ou min_conf > 0.8, aucune règle d'association forte n'est générée. Ces seuils plus élevés imposent des critères plus stricts, réduisant ainsi la quantité de règles extraites, mais potentiellement augmentant leur qualité.

En outre, nous avons constaté que le nombre maximal de règles générées atteint 35, indiquant une limite à la complexité du modèle d'association dans notre configuration expérimentale.

Ces observations suggèrent l'importance de choisir judicieusement les valeurs de minsup et minconf en fonction des objectifs spécifiques de l'analyse, et elles fournissent une base pour des investigations plus approfondies dans le cadre de notre étude des règles d'association.

Recommandations basées sur l'algorithme Apriori

Dans le cadre de notre étude, nous explorons l'utilisation de l'algorithme Apriori pour générer des recommandations personnalisées dans le contexte spécifique de l'agriculture. L'objectif est d'améliorer les pratiques agricoles en suggérant des cultures et des fertilisants appropriés en fonction des conditions environnementales et des besoins spécifiques des utilisateurs

Lorsqu'un agriculteur insère de nouvelles données, telles que la température, l'humidité, les précipitations, le type de sol, la culture prévue, et le fertilisant potentiel, notre système utilise l'algorithme Apriori pour générer des recommandations. Par exemple, il peut suggérer des cultures adaptées à ces conditions spécifiques ainsi que des fertilisants qui ont montré des associations fréquentes dans des situations similaires.

L'ajout de la nouvelle ligne d'utilisateur, comprenant les informations sur la température, l'humidité, les précipitations, le type de sol, la culture envisagée, et le fertilisant potentiel, permet de fournir des recommandations encore plus personnalisées pour ce même utilisateur.
S'il insère ces données :
{ temperature=25.5,humidity=60.0,rainfall=1.2, soil='Loam', crop='Coconut',fertilizer='DAP'} , le système lui retournera des graines et des fertilisateurs qu'il pourra utiliser pour optimiser les pratiques agricoles en suggérant des cultures et des fertilisants adaptés à ses conditions spécifique. Dans notre cas, le système lui recommande 'Coconut' et 'MOP' .

Cette approche s'inscrit dans notre volonté d'exploiter l'intelligence des données pour améliorer la productivité agricole et la durabilité environnementale.

Conclusion

En conclusion de notre étude approfondie sur l'analyse et le prétraitement des données, ainsi que des visualisations significatives des données temporelles, nous avons réalisé une exploration exhaustive de trois jeux de données distincts : la fertilité du sol, les données temporelles liées au COVID-19 aux États-Unis, et les données climatiques et agricoles. Notre objectif principal était d'optimiser l'application de la classification à nos deux versions du premier dataset afin d'en étudier les effets.

En somme, cette étude a fourni une compréhension approfondie de diverses techniques d'analyse et de prétraitement des données, ainsi que des visualisations significatives des données temporelles. L'application de l'algorithme Apriori a ajouté une dimension de recommandations utiles. Notre prochaine étape consistera à appliquer la classification à nos deux versions du dataset portant sur la fertilité du sol, et nous anticipons avec enthousiasme les résultats qui en découleront. Cette approche analytique et pratique renforcera notre capacité à prendre des décisions éclairées dans le domaine de l'optimisation agricole.

Références

[1] Soil Nitrate Measurement for Determination of Plant-Available Nitrogen - HORIBA., consulté le 24 Novembre

[2] -Critical limits of available P in soils for different crops and soils | Download Table., consulté le 24 Novembre

[3] Image: Range-and-mean-values-of-some-properties-of-the-studied-soils, The forms and availability to plants of soil potassium as related to mineralogy for upland Oxisols and Ultisols from Thailand, https://www.researchgate.net/profile/Timtong-Darunsontaya/publication/241126981/figure/tbl2/AS:783464978722818@1563804015244/Range-and-mean-values-of-some-properties-of-the-studied-soils.png, consulté le 24 Novembre

[4] Image: Relative-levels-of-soil-test-potassium, Soil Test Interpretations and Fertilizer Management for Lawns, Turf, Gardens, and Landscape Plants, https://www.researchgate.net/publication/265451223/figure/tbl3/AS:669564882915333@1536648116134/Relative-levels-of-soil-test-potassium.png, consulté le 24 Novembre

[5] Relative levels of soil test potassium | Download Table, consulté le 24 Novembre

[6] Standard operating procedure for soil pH determination, consulté le 24 Novembre

[7] What can electrical conductivity tell us about our soil?., consulté le 24 Novembre

[8] What is soil organic carbon? | Agriculture and Food., consulté le 24 Novembre

[9] How Much Elemental Sulphur To Apply?, consulté le 24 Novembre

[10] Establishment of Critical Limits of Zinc in Soils Using Multi-extractants for Paddy Crop Grown in Central India, consulté le 24 Novembre

[11]https://kiran.nic.in/pdf/publications/2017/Boron_Nutrition_in_Soil_System_and_Management_Strategy.pdf?fbclid=IwAR10mHzLeRo5YBy73rlFmySEPhKUG28TfJyW9Rk1NVuF_OwOR3RmIcfWqvQ, consulté le 24 Novembre

[12]Ecological Soil Screening Level for Iron, U. S. Environmental Protection Agency Office of Solid Waste and Emergency Response 1200 Pennsylvania Avenue, N.W. Washington, DC 20460 November 2003, https://rais.ornl.gov/documents/eco-ssl_iron.pdf?fbclid=IwAR3DHmWvWjJn1JvUiA1MKhNrc4ShnPHl2kMa3WjbifQ1XwKfqY0YBIzoqfc, consulté le 24 Novembre

[13] MANGANESE, AGRITOPIC January 2020 https://www.incitecpivotfertilisers.com.au/~/media/Files/IPF/Documents/Agritopics/24%20Manganese%20Agritopic.pdf?fbclid=IwAR3RIsWbxx_UXl65q7W8erEgiBVJUIGv5j2taNDPf-nn0wkSGmoPn8EQ5_c, consulté le 24 Novembre

[14] Ch 3. Amount of Organic Matter in Soils - SARE, consulté le 24 Novembre

[15] Algorithme APriori — Wikipédia, consulté le 25 Novembre