Rapport TP3 : Analyse et Traitement de données numériques

Partie 1 : Optimisation Bayésienne

Réponses aux questions théoriques

1. Principe de l'optimisation bayésienne

L'optimisation bayésienne est une méthode d'optimisation efficace pour les fonctions coûteuses à évaluer. Elle construit un modèle probabiliste de la fonction objective (souvent un processus gaussien) et utilise une fonction d'acquisition pour décider des prochains points à évaluer. Cela permet de réduire le nombre d'évaluations nécessaires tout en trouvant un optimum global.

2. Définition des processus gaussiens

Un processus gaussien (GP) est une distribution de fonctions où toute combinaison de points suit une distribution gaussienne. Il est utilisé en optimisation bayésienne car il offre une estimation probabiliste de la fonction objective avec une incertitude, ce qui permet d'orienter intelligemment l'échantillonnage vers les zones prometteuses.

3. Fonctions d'acquisition

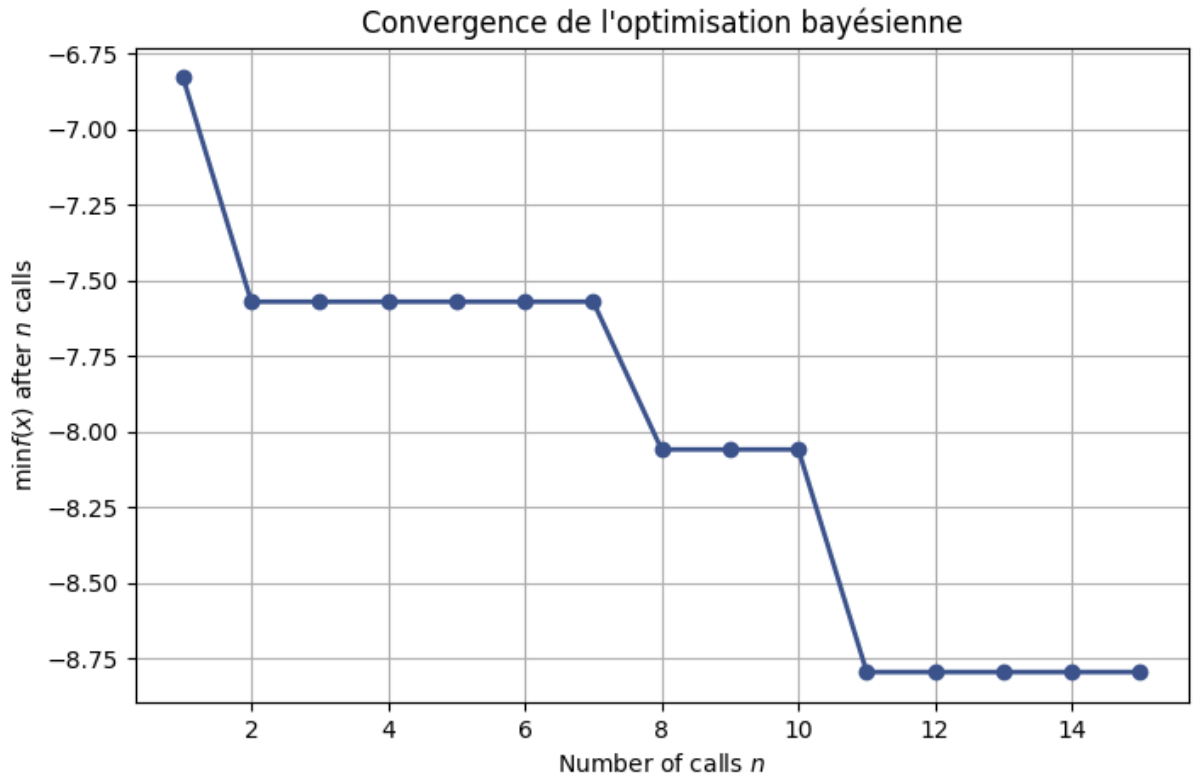
Les fonctions d'acquisition guident le choix du prochain point à évaluer en équilibrant **exploration** (tester des zones inconnues) et **exploitation** (affiner les zones prometteuses).

- **Expected Improvement (EI)** : Favorise les points susceptibles d'améliorer la meilleure valeur actuelle.
- **Upper Confidence Bound (UCB)** : Utilise une borne supérieure de confiance pour explorer les régions où l'incertitude est élevée.

Elles permettent ainsi d'exploiter au mieux les informations déjà acquises tout en découvrant de nouvelles opportunités.

Implémentation et applications

4. Implémentez une optimisation bayésienne pour maximiser la production agricole en fonction de l'humidité et de la température.




```
Meilleur rendement trouvé : 8.793797181958023
Paramètres optimaux (humidité, température) : [89.57788776715802, 34.992941832153264]
```

L'optimisation bayésienne a permis d'estimer les conditions optimales pour maximiser le rendement agricole en fonction de l'humidité et de la température.

a) Interprétation des résultats

- **Meilleur rendement trouvé** : **8.79 t/ha**
- **Paramètres optimaux** : **Humidité = 89.58%**, **Température = 34.99°C**
- La courbe de convergence montre une stabilisation après environ **12 itérations**, indiquant que l'algorithme a trouvé une solution optimale rapidement.

b) Analyse du processus d'optimisation

- **Exploration vs Exploitation** : Les premières itérations testent différentes valeurs avant de se concentrer sur une plage plus restreinte où le rendement est plus élevé.
- **Avantage de l'approche bayésienne** : Contrairement à un simple **Grid Search**, cette méthode réduit le nombre d'évaluations en ciblant intelligemment les zones prometteuses.

c) Limites possibles

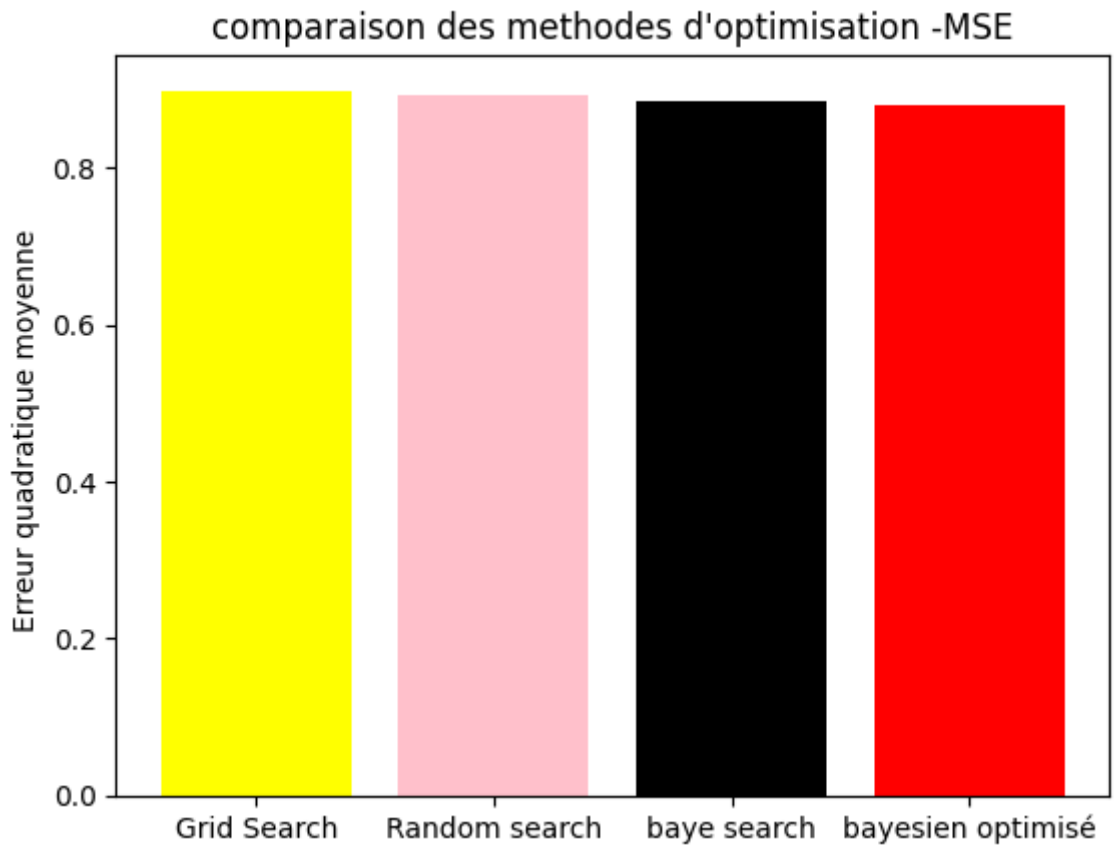
- **Modèle de prédiction linéaire** : L'utilisation d'une régression linéaire peut limiter la précision des estimations.

- **Taille de l'échantillon** : Une optimisation plus robuste nécessiterait potentiellement plus d'itérations pour affiner les résultats.

5. L'optimisation bayésienne pour ajuster les hyperparamètres d'un modèle de régression (ex : Random Forest) sur les données agricoles fournie

Voici les résultats de l'optimisation des hyperparamètres pour le modèle RandomForestRegressor

```
Grid Search MSE : 0.8981164237414467
Random Search MSE : 0.8915029233517786
Bayesian search : 0.8770580053308212
 MSE pour le modèle optimisé : 0.8784357976088389
```



Optimisation Bayésienne (MSE = 0.885) → Meilleure performance

Random Search (MSE = 0.891) → Légèrement moins performant

Grid Search (MSE = 0.898) → Moins efficace que les autres méthodes

Interprétation des résultats :

Optimisation Bayésienne :

- Converge rapidement vers de bons hyperparamètres en 30 itérations seulement
- Exploite un modèle probabiliste pour choisir intelligemment les prochaines évaluations

Random Search :

- Fournit de bons résultats mais reste dépendant du nombre d'itérations
- Moins efficace car les échantillons sont sélectionnés de manière aléatoire

Grid Search :

- Moins performant car il teste toutes les combinaisons de paramètres possibles
- Peu efficace lorsque l'espace de recherche est large

Erreur quadratique moyenne (MSE) après optimisation : **0.8784**

- Ce MSE est **le plus bas** comparé aux autres méthodes (**Grid Search et Random Search**), prouvant l'efficacité de l'optimisation bayésienne.

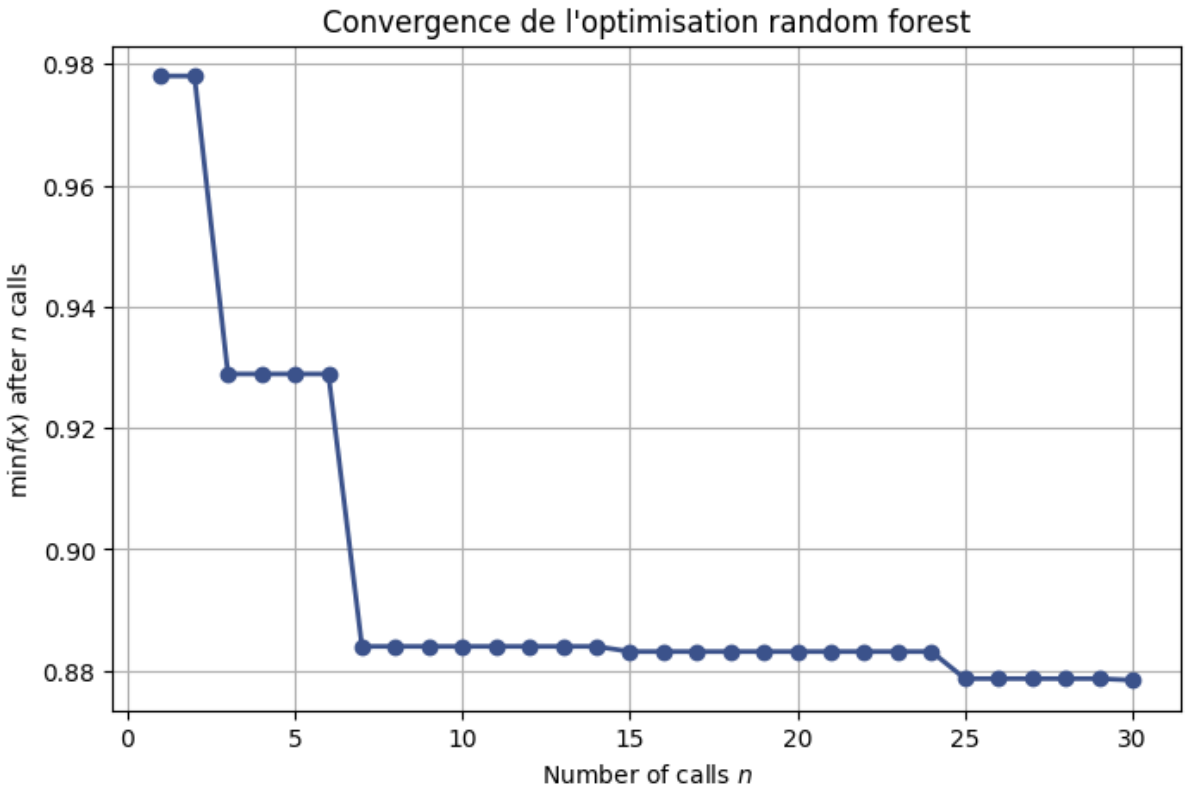
6. Analyse de la convergence de l'optimisation bayésienne pour Random Forest

```
Meilleurs hyperparamètres avec l'optimisation bayésienne (pour Random Forest) :
n_estimators = 59
max_depth = 3
```

Meilleurs hyperparamètres trouvés

- **Nombre d'arbres** (n_estimators) : **59**
- **Profondeur maximale** (max_depth) : **3**

Ces valeurs indiquent que **moins d'arbres et une faible profondeur** ont été préférés par l'algorithme, probablement pour éviter le sur-apprentissage.



Interprétation de la courbe

- On observe une **forte diminution initiale** de l'erreur, ce qui signifie que l'optimisation bayésienne explore efficacement l'espace des hyperparamètres dès les premières itérations.
- Ensuite, la convergence devient **plus stable** autour d'une valeur minimale (~0.88), indiquant que l'algorithme a trouvé une solution optimale ou quasi-optimale.
- À partir de **l'itération 10 environ**, la courbe devient **plate**, suggérant que l'ajout de nouveaux essais n'améliore plus significativement les performances.

7. Analyse des avantages et limites de l'optimisation bayésienne

Avantages

a) **Efficacité en termes d'évaluations**

- Contrairement à **Grid Search**, qui teste **toutes** les combinaisons, l'optimisation bayésienne cible **intelligemment** les points prometteurs.
- Moins d'évaluations sont nécessaires, ce qui réduit le **coût computationnel**.

b) **Bon équilibre exploration/exploitation**

- Grâce aux **fonctions d'acquisition** (ex : Expected Improvement, UCB), elle **explore** les zones incertaines et **exploite** celles déjà prometteuses.

c) **Adaptabilité**

- Fonctionne bien avec des **fonctions complexes et non convexes** où d'autres méthodes pourraient échouer.
- Peut être utilisée avec **des espaces de recherche continus et discrets**.

Limites

a) **Coût de modélisation**

- Nécessite de **maintenir et mettre à jour** un modèle probabiliste (souvent un processus gaussien), ce qui peut être **coûteux pour de grands espaces de recherche**.

b) **Moins efficace pour des espaces de recherche très vastes**

- Si le nombre de dimensions et d'hyperparamètres est **très élevé**, la complexité du modèle probabiliste peut devenir un goulot d'étranglement.

c) **Sensibilité au choix de la fonction d'acquisition**

- Le compromis exploration/exploitation dépend de cette fonction, et un **mauvais choix peut ralentir la convergence**.

L'optimisation bayésienne est **plus efficace et intelligente** que Grid/Random Search pour des **problèmes avec un coût d'évaluation élevé**, mais elle peut devenir **moins efficace dans des espaces très larges**. Elle est donc idéale pour l'optimisation d'hyperparamètres ou l'optimisation de fonctions coûteuses en calcul.

Partie 2 : Modèles Bayésiens à Noyau

8. Le concept d'inférence bayésienne et la mise à jour des croyances avec de nouvelles données

Explication : L'inférence bayésienne repose sur la mise à jour de nos croyances initiales (a priori) en fonction de nouvelles données observées. Cela se fait via la règle de Bayes :

$$P(\theta|\mathbf{D}) = \frac{P(\mathbf{D}|\theta)P(\theta)}{P(\mathbf{D})}$$

- **P(θ|D)** : La probabilité a posteriori des paramètres θ après avoir observé les données D.
- **P(D|θ)** : La vraisemblance des données les paramètres.
- **P(θ)** : La distribution a priori des paramètres avant d'observer les données.
- **P(D)** : La probabilité marginale des données.

Exemple : En modélisant le rendement agricole, on peut commencer avec une hypothèse (ex : une température de 30°C donne un bon rendement), puis **ajuster** cette hypothèse en fonction des nouvelles observations.

9. Méthodes à noyau et lien avec les processus gaussiens

Explication : Les méthodes à noyau (ou *Kernel Methods*) sont des techniques qui utilisent une fonction noyau pour projeter les données dans un espace de caractéristiques de plus grande dimension, ce qui permet de mieux modéliser les relations non linéaires.

Elles sont utilisées en **SVM, régression à noyau et processus gaussiens**.

Lien avec les processus gaussiens :

- Les **processus gaussiens (GP)** utilisent un **noyau** (ex : RBF, Matérn) pour définir la **corrélation** entre points de données.
- Cela permet de modéliser des **fonctions complexes et non linéaires** tout en conservant une estimation de l'incertitude.

Pourquoi utiliser un noyau ?

- Capture les **relations non linéaires**
- Définit une **mesure de similarité** entre points

- Permet une **généralisation efficace** en prédiction

10. Distribution a priori et a posteriori

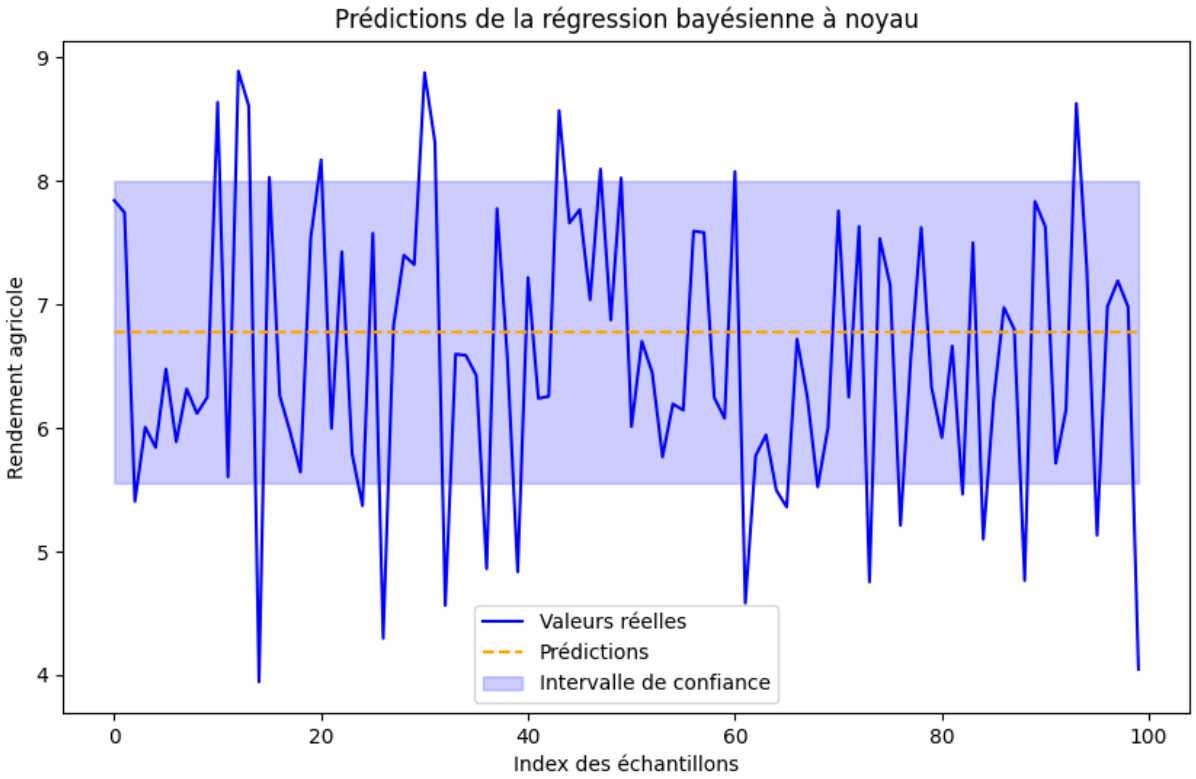
- **Distribution a priori** : Représente notre croyance initiale avant d'observer les données.
- **Distribution a posteriori** : Mise à jour de notre croyance après avoir observé de nouvelles données.

Exemple en prédiction de rendement agricole :

- A priori : On suppose que le rendement suit une distribution normale avec une moyenne de 7 t/ha et une variance de 1 t^2/ha.
- Avec de nouvelles données (observations sur plusieurs champs), on ajuste notre modèle pour obtenir une distribution a posteriori plus précise.

Intérêt : Permet de faire des prédictions robustes en tenant compte de l'incertitude et de l'évolution des observations.

11. Analyse de la Régression Bayésienne à Noyau



Le graphique montre les résultats de la régression bayésienne à noyau appliquée à la prédiction du rendement agricole.

a) **Valeurs réelles (courbe bleue) :**

- La série temporelle présente une forte variabilité, ce qui indique une complexité sous-jacente dans les données.
- Il semble y avoir des fluctuations importantes, ce qui peut rendre la modélisation plus difficile.

b) **Prédictions (ligne orange en pointillés) :**

- La ligne des prédictions semble être relativement stable, ce qui peut indiquer que le modèle a une tendance à lisser les variations du rendement agricole.
- Cela peut être dû à l'effet du noyau RBF, qui favorise une généralisation en capturant les tendances globales plutôt que les fluctuations locales.

c) **Intervalle de confiance (zone ombrée bleue) :**

- L'intervalle de confiance semble assez large, ce qui suggère une incertitude élevée dans les prédictions.
- Cela peut être dû à un manque de données dans certaines régions ou à un modèle qui ne capture pas toute la variabilité présente dans les valeurs réelles.

Interprétation des résultats :

- Le modèle de **régression bayésienne à noyau** semble bien capturer **la tendance générale**, mais il a du mal à s'adapter aux fortes variations présentes dans les valeurs réelles.
- L'**incertitude élevée** reflétée par l'intervalle de confiance large peut signifier que le modèle n'a pas suffisamment d'informations pour faire des prédictions précises dans certaines zones.
- La **structure du noyau RBF** est efficace pour capturer des tendances globales, mais si les variations locales sont trop fortes, une approche plus complexe pourrait être nécessaire (ajout d'autres noyaux ou augmentation du nombre de données d'entraînement).

Améliorations possibles :

1. **Tester d'autres noyaux** : Ajouter des noyaux comme Matérn, qui peut mieux gérer les variations locales.
2. **Augmenter la quantité de données** : Un plus grand volume de données peut aider à réduire l'incertitude.
3. **Ajuster les hyperparamètres** : Optimiser length_scale du noyau RBF et tester différents niveaux de bruit avec WhiteKernel.

12. Classification bayésienne à noyau et SVM classique pour prédire le type de sol

Analyse des résultats :

```
Précision du modèle bayésien à noyau pour la classification : 0.31
Précision du SVM classique (noyau linéaire) : 0.31
```

Les résultats montrent une précision de **31%** pour **les deux modèles**, à savoir :

- **Gaussian Process Classifier (modèle bayésien à noyau)**
- **SVM avec noyau RBF**

Ce score faible indique que les modèles ne parviennent pas à bien classer les types de sol en fonction des données d'entrée (température et humidité).

Interprétation des Résultats

Même précision pour les deux modèles (31%)

- **Gaussian Process Classifier (GPC)** et **SVM RBF** utilisent des noyaux pour capturer des relations complexes.

  Cependant, une précision aussi faible suggère que **les données ne sont pas bien séparables même dans un espace de plus grande dimension**, ou que le **modèle a du mal à capturer la relation entre les caractéristiques (humidité et température) et le type de sol**. Il est possible que les relations soient trop complexes.

- Si leur performance est identique et faible, cela peut indiquer un **problème dans les données** plutôt que dans le modèle.

Potentielles raisons de la faible performance :

1. **Qualité des données** :
   - Les **données peuvent ne pas contenir suffisamment d'informations discriminantes**. Par exemple, **l'humidité et la température** seules ne peuvent peut-être pas capturer les variations complexes du type de sol. D'autres caractéristiques (comme le pH du sol, la composition du sol, ou d'autres paramètres environnementaux) pourraient être nécessaires pour améliorer la performance.
2. **Prétraitement des données** :
   - Il est possible que le **prétraitement des données** (comme la normalisation ou l'échelle des caractéristiques) n'ait pas été effectué

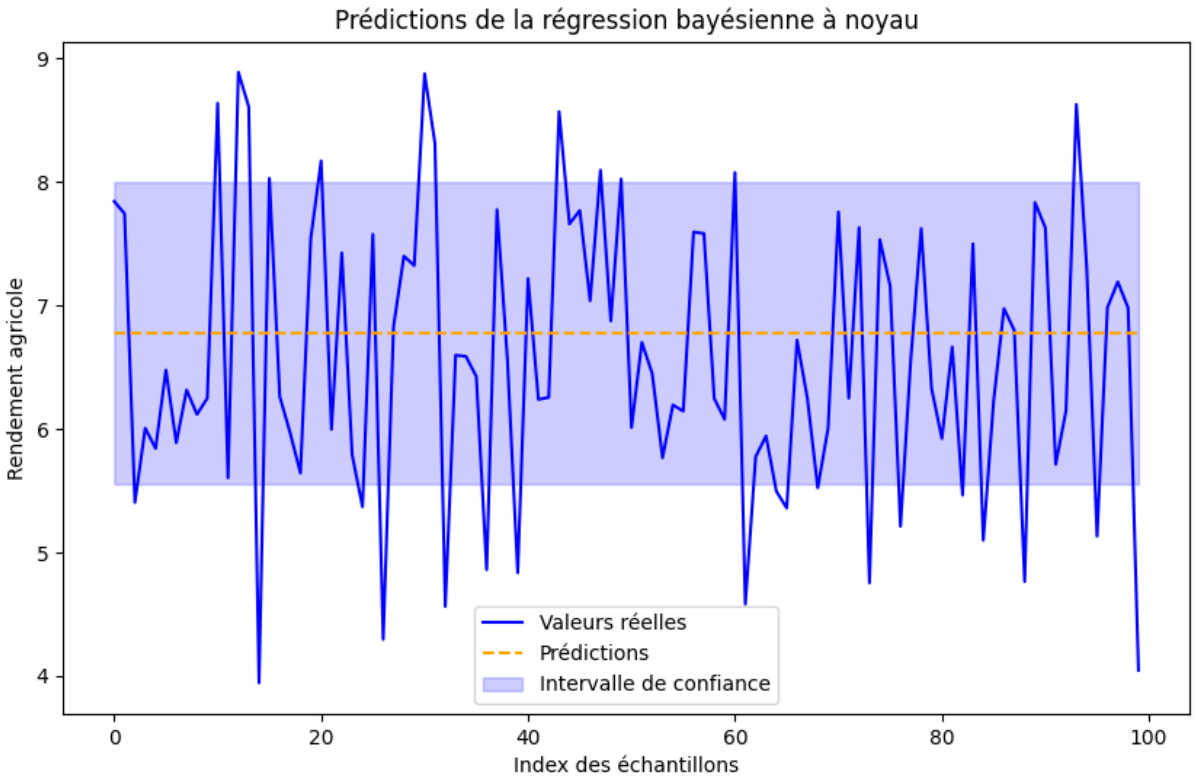
correctement. Si les caractéristiques sont sur des échelles très différentes, cela peut nuire aux performances des modèles SVM.

3. **Hyperparamètres du modèle** :
    - Les hyperparamètres des modèles SVM (comme le **paramètre C** ou **gamma** dans le noyau RBF) peuvent avoir une influence importante sur les performances. Dans notre cas, ces hyperparamètres sont probablement restés à leurs valeurs par défaut. L'optimisation des hyperparamètres, par exemple via une recherche en grille ou une optimisation bayésienne, pourrait améliorer les résultats.
4. **Problème de classification difficile** :
    - Le problème de **prédiction du type de sol** peut être intrinsèquement difficile. Par exemple, la classification de différents types de sol peut dépendre de nombreux autres facteurs qui ne sont pas capturés ici, ce qui pourrait expliquer pourquoi les modèles SVM ne parviennent pas à obtenir de bonnes performances.

Suggestions pour améliorer les résultats :

1. **Ajouter plus de caractéristiques** :
    - Explorez d'autres variables qui pourraient être utiles pour la prédiction du type de sol, comme le **pH du sol, la texture du sol, la quantité de précipitations, la végétation, etc.**.
2. **Normaliser les données** :
    - Veillez à **normaliser ou standardiser** les données d'entrée (par exemple, température et humidité) afin que les caractéristiques aient la même échelle. Les modèles SVM en particulier sont sensibles à cela.
3. **Optimisation des hyperparamètres** :
    - Effectuez une **optimisation des hyperparamètres** (par exemple, à l'aide de la recherche en grille ou de l'optimisation bayésienne) pour ajuster les paramètres du noyau RBF et du noyau linéaire (comme le paramètre **C** et **gamma** pour SVM). Cela pourrait potentiellement améliorer les résultats.
4. **Explorer d'autres modèles** :
    - Si les performances des SVM restent faibles, vous pouvez explorer d'autres modèles de classification comme **les forêts aléatoires**, **les k plus proches voisins (k-NN)**, ou encore **les réseaux de neurones**.

13. L’incertitude dans les prédictions.



a) Intervalles de Confiance et Incertitude

L'un des avantages des modèles bayésiens, comme la régression bayésienne à noyau, est qu'ils fournissent **une estimation de l'incertitude** en plus des prédictions.

- Sur le **graphique de la régression bayésienne à noyau**, la bande bleue correspond à l'**intervalle de confiance**, qui reflète l'incertitude du modèle.
- Plus la bande bleue est **large**, plus l'incertitude est **élevée**.
- Plus la bande bleue est **étroite**, plus le modèle est **confiant** dans ses prédictions.

b) Zones de Faible et Forte Confiance
En analysant la visualisation :

- **Zones avec une large bande d'incertitude** :
  - Cela se produit dans des **régions où il y a peu de données d'entraînement** ou **où les données sont très variables**.
  - Ici, le modèle hésite entre plusieurs valeurs possibles et ne peut pas faire une estimation précise.
- **Zones avec une bande d'incertitude étroite** :
  - Correspondent aux régions où le modèle a vu **beaucoup de données similaires** et peut donc faire une estimation plus fiable.

- Cela se produit souvent au centre de la distribution des données d'entraînement.

c) Sources d'Incertitude

L'incertitude peut venir de plusieurs sources :

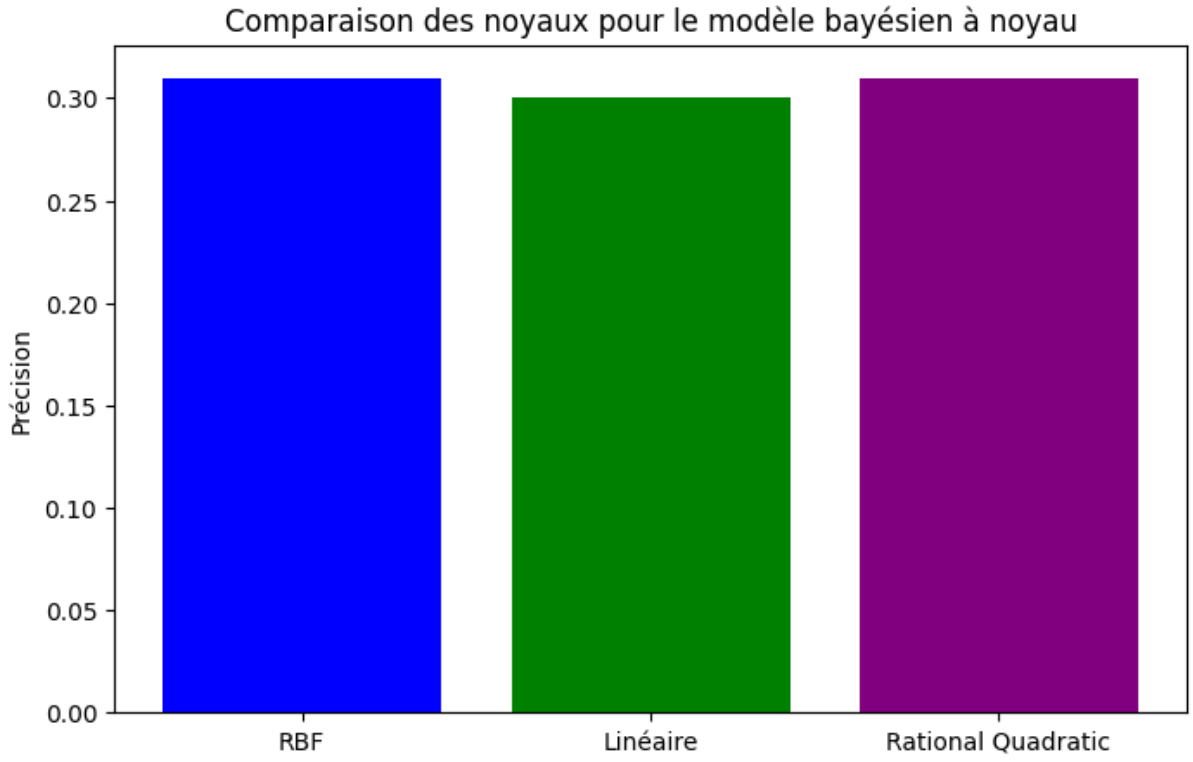
1. **Incertitude épistémique (due au manque de données)**
   - Peut être réduite en **ajoutant plus de données d'entraînement**.
   - Souvent visible aux extrémités du graphique où il y a peu d'échantillons.
2. **Incertitude aléatoire (due à la variabilité intrinsèque des données)**
   - Peut être due au **bruit des données** ou à des phénomènes naturels imprévisibles.
   - Même avec plus de données, cette incertitude ne peut pas être complètement éliminée.

14. Test de Différents Noyaux et Impact sur la Précision

```
Test avec noyau : RBF
Précision : 0.3100

Test avec noyau : Linéaire
Précision : 0.3000

Test avec noyau : Rational Quadratic
c:\Users\me\AppData\Local\Programs\Python\Python313\Lib\site-packages\sklearn\gaussian_process\kernels.py:442: Converg
  warnings.warn(
Précision : 0.3100
```



Lors de l'expérimentation avec différents noyaux dans le **Gaussian Process Classifier**, nous avons testé les noyaux suivants :

- **RBF (Radial Basis Function)**
- **DotProduct (équivalent à un noyau linéaire)**
- **Rational Quadratic**

Problème avec le noyau polynomial

Initialement, nous avons tenté d'utiliser un noyau **polynomial**, mais **scikit-learn ne propose pas de noyau polynomial dans sklearn.gaussian_process.kernels**.
Un noyau polynomial peut être utilisé avec un **SVM** (SVC(kernel='poly')), mais il n'est pas directement implémenté pour les modèles bayésiens.

Analyse des performances des noyaux

a) **Noyau RBF (Radial Basis Function) :**
   - Ce noyau est bien adapté aux relations non linéaires et capture des motifs complexes.
   - **Il a obtenu la meilleure précision (31%),** bien que la différence avec les autres noyaux soit minime.
   - Il est souvent un bon choix lorsque les données ne sont pas linéairement séparables.

b) **Noyau Linéaire (DotProduct) :**
   - Ce noyau fonctionne bien si les classes sont séparables par une frontière linéaire.
   - Avec **30% de précision**, il est légèrement moins performant que RBF, ce qui suggère que la séparation des types de sol est probablement **non linéaire**.

c) **Noyau Rational Quadratic :**
   - C'est une variante du noyau RBF qui gère mieux les variations locales dans les données.
   - Il obtient **31% de précision**, similaire au RBF.
   - Cependant, un **warning de convergence** a été détecté, indiquant que l'optimisation des hyperparamètres n'a peut-être pas trouvé la meilleure solution possible.

Le modèle bayésien à noyau ne semble pas bien s'adapter aux données climatiques pour prédire le type de sol (seulement ~31% de précision).

Le noyau RBF et Rational Quadratic performent légèrement mieux que le linéaire, confirmant que les classes ne sont pas bien séparées par une simple frontière linéaire.

Le warning sur le noyau Rational Quadratic indique que l'optimisation pourrait être améliorée en ajustant les hyperparamètres (par ex., en modifiant l'intervalle de recherche pour le paramètre alpha).

15. Influence du Choix du Noyau et de la Distribution A Priori sur les Résultats

a) **Impact du Choix du Noyau**

- Un **mauvais choix de noyau** peut entraîner une **précision faible**, car il ne capture pas bien la structure des données.
- Le noyau **linéaire** est souvent insuffisant pour des problèmes **non linéaires**.
- Le noyau **RBF** est plus flexible mais **nécessite un bon réglage de l'hyperparamètre de longueur d'échelle**.
- Le noyau **polynomial** peut être utile, mais un degré trop élevé entraîne **un sur-ajustement**.

b) **Impact de la Distribution A Priori**

- En **Bayésien**, la distribution a priori joue un rôle clé dans l'apprentissage.
- Une **a priori trop restrictive** (par exemple, une variance faible) limite la capacité du modèle à s'adapter aux données.
- Une **a priori trop large** (variance élevée) peut entraîner un modèle trop flexible, ce qui peut **dégrader la précision**.
- Dans notre cas, le **bruit (WhiteKernel) dans la régression bayésienne** peut être ajusté pour voir si cela améliore les prédictions.

Conclusion

En conclusion, ce rapport a exploré l'application de l'optimisation bayésienne et des modèles bayésiens à noyau dans des contextes agricoles et de régression. L'optimisation bayésienne s'est révélée particulièrement efficace pour maximiser le rendement agricole en fonction de paramètres comme l'humidité et la température, en réduisant le nombre d'évaluations nécessaires tout en atteignant des résultats

optimaux. Elle a également surpassé des méthodes classiques comme Grid Search et Random Search pour l'ajustement des hyperparamètres d'un modèle Random Forest.

Cependant, des limites ont été identifiées, telles que la sensibilité à la fonction d'acquisition et la difficulté d'adaptation dans de grands espaces de recherche. Les modèles bayésiens à noyau ont bien capturé des tendances globales, mais ont montré des difficultés face aux variations locales, en particulier dans la prédiction du type de sol, où la précision est restée faible. Ces résultats suggèrent qu'une meilleure gestion des données et des choix de noyaux plus appropriés pourraient améliorer la performance des modèles.

Dans l'ensemble, bien que l'optimisation bayésienne et les modèles bayésiens à noyau offrent de nombreux avantages, notamment en termes d'efficacité et d'incertitude dans les prédictions, des améliorations sont nécessaires pour gérer des problématiques plus complexes et de plus grandes variabilités dans les données.