Rapport de Modélisation Prédictive des Températures Océanographiques

Ayoub LAMCHICHI - Matthias TRUPIN

1. Présentation du Dataset

1.1 Source et Description

Ce projet utilise les données océanographiques du programme CalCOFI (California Cooperative Oceanic Fisheries Investigations), qui collecte des données sur les caractéristiques physiques et chimiques de l'océan depuis plus de 70 ans. Ces données sont essentielles pour comprendre les écosystèmes marins et les changements climatiques.

1.2 Aperçu des Données

Le jeu de données contient des mesures océanographiques incluant:

- **Température**: Température de l'eau en degrés Celsius
- **Salinité**: Concentration en sel de l'eau
- **Profondeur**: Profondeur de la mesure en mètres
- **Oxygène**: Concentration en oxygène dissous
- **Densité**: Densité de l'eau

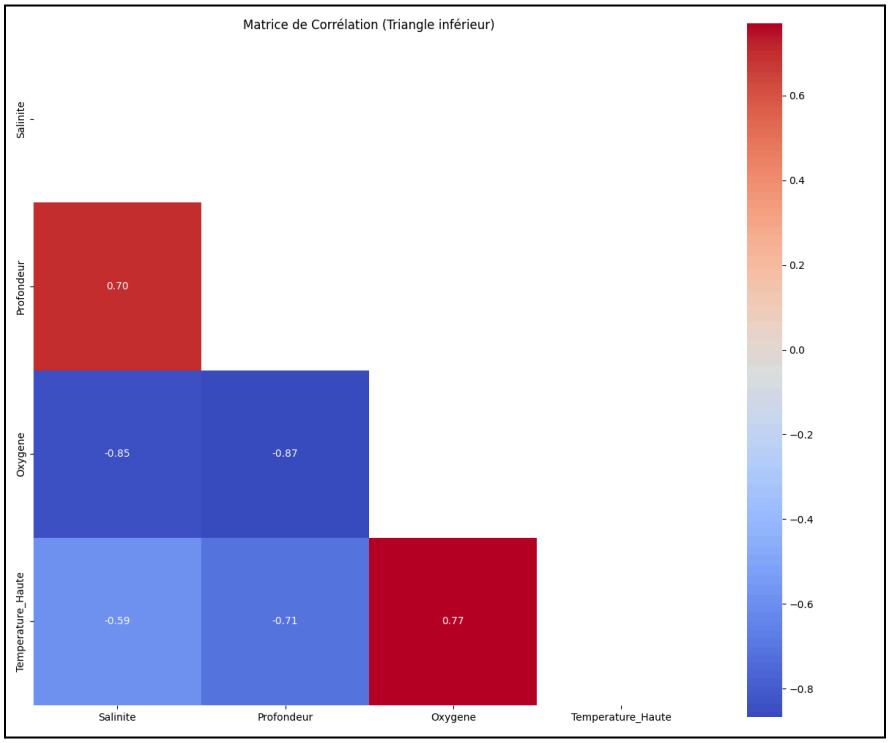
Taille du dataset: 621740 observations

1.3 Statistiques Descriptives

	Salinité	Profondeur	Oxygène	Densité	Température
count	621740	621740	621740	621740	621740
moyenne	33.79	168.44	3.55	25.73	11.21
ecart-type	0.43	158.44	2.02	0.94	3.85
minimum	32.44	0.00	0.00	22.71	2.88
25%	33.46	45.00	1.68	24.91	8.18
50% (médiane)	33.79	119.00	3.70	25.88	10.41
75%	34.13	250.00	5.56	26.56	14.13
maximum	35.15	676.00	11.13	27.42	23.25

1.4 Analyse des Corrélations

Pour identifier les variables qui ont le plus de poids dans la détermination de la température, nous avons établi une matrice de corrélation qui révèle les influences entre les variables:



On observe sur cette matrice :

- Forte corrélation négative entre Oxygène et Profondeur (-0.87)
- Corrélation négative entre Oxygène et Salinité (-0.85)
- Forte corrélation positive entre Température_Haute et Oxygène (0.77)
- Corrélation négative entre Température_Haute et Profondeur (-0.71)
- Corrélation négative entre Température_Haute et Salinité (-0.59)
- Corrélation positive entre Salinité et Profondeur (0.70)

Ces corrélations sont cohérentes avec les principes océanographiques connus:

- L'eau froide des profondeurs contient généralement plus d'oxygène dissous
  - *“Ainsi, l'eau froide peut contenir une concentration plus élevée d'oxygène dissous que l'eau chaude”* - *Source*

- La salinité tend à augmenter avec la profondeur
  - *“Le seuil de concentration massique augmente [...] Il augmente aussi avec la pression”* *Source*

- Les eaux plus chaudes se trouvent généralement naturellement en surface

2. Préparation des Données

2.1 Approches Utilisées

Dans cette étude, nous avons adopté deux approches complémentaires:

1. **Classification binaire**: Prédire si la température est haute ou basse
2. **Régression**: Prédire la valeur exacte de la température

2.2 Traitement pour la Classification

Pour la classification binaire, nous avons créé une variable cible binaire **Temperature_Haute** qui est

- soit au-dessus de la médiane (10.41°C)
- soit en dessous de cette médiane.

Nous avons ensuite testé deux ensembles de variables explicatives:

- Avec densité: Salinité, Profondeur, Oxygène, Densité
- Sans densité: Salinité, Profondeur, Oxygène

2.3 Traitement pour la Régression

Après avoir utilisé la classification, nous avons utilisé la régression qui assimile mieux la complexité du sujet, nous avons ainsi :

- Utilisé la température directement comme variable cible continue
- Utilisé comme variables explicatives: la Salinité, la Profondeur et l'Oxygène
- Évité d'utiliser la densité pour éviter toute circularité.

2.4 Prétraitement des Données

Pour les deux approches, nous avons appliqué les traitements communs suivants:

- Division train/test (80% / 20%)
- Standardisation des variables numériques (moyenne = 0, écart-type = 1)
- Vérification de l'équilibre des classes pour la classification (distribution des classes: {0: 0.50, 1: 0.50})
- Taille de l'ensemble d'entraînement: 497.392 observations
- Taille de l'ensemble de test: 124.348 observations

3. Modélisation et Résultats

3.1 Classification Binaire

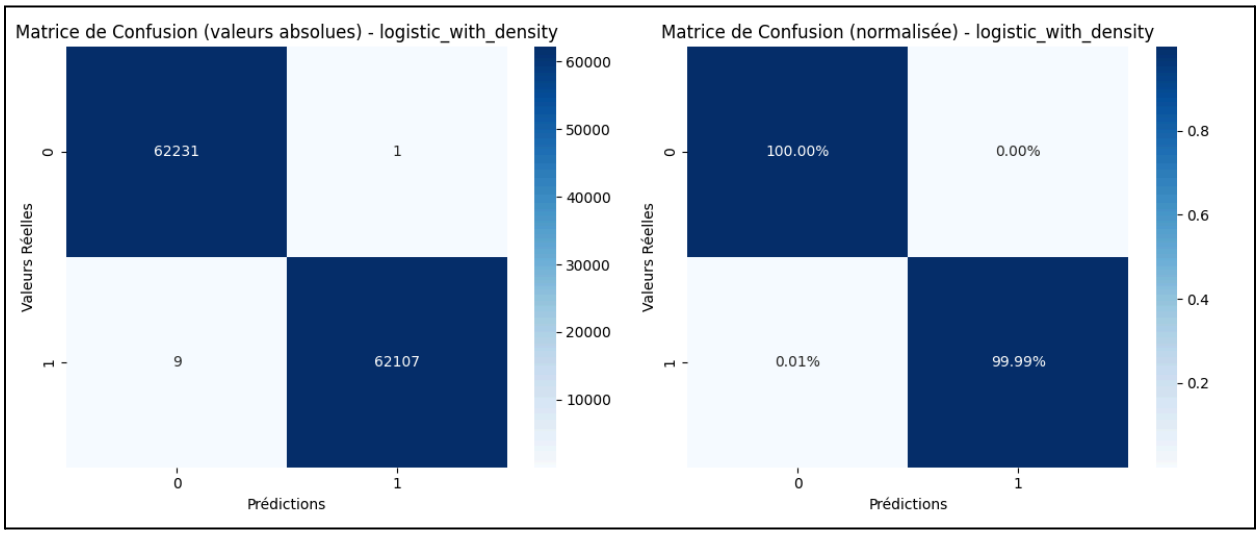
3.1.1 Modèle avec densité

Le deuxième entraînement concerné, nous avons ensuite entraîné une régression logistique en incluant la densité comme variable explicative.

Voici les **performances** mesurées avec la densité au bout de 5000 itérations:

- Accuracy: 0.9999
- Precision: 1.0000
- Recall: 0.9999
- F1 Score: 0.9999
- AUC-ROC: 1.0000

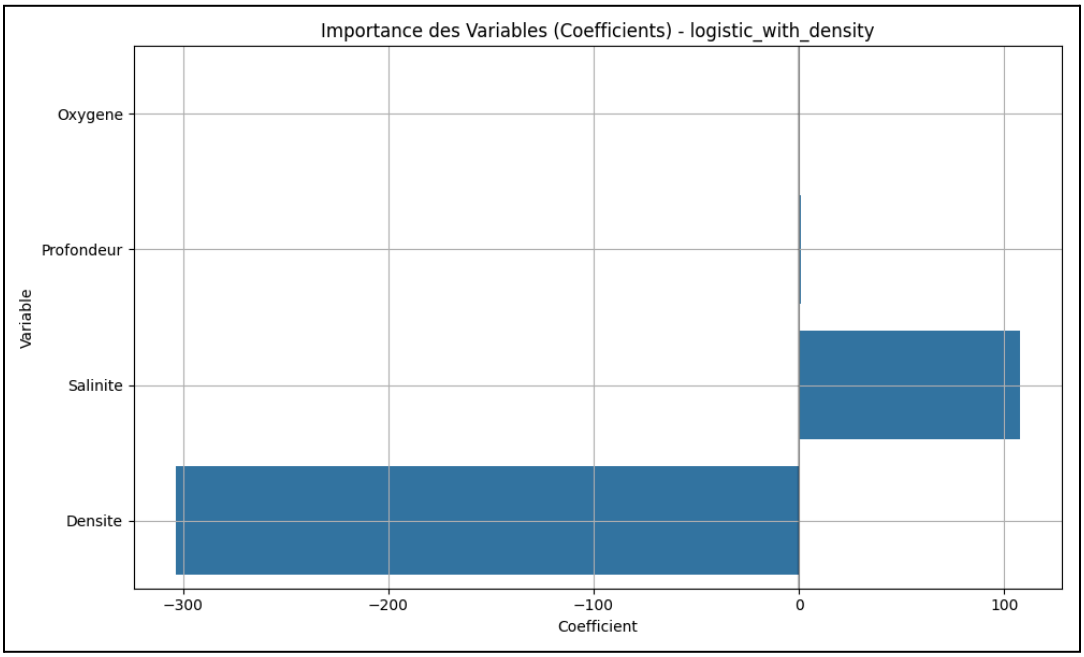
Matrice de confusion :



Voici les observations obtenus à la lecture de cette matrice :

- Vrai Négatif: 62231 (100.00%)
- Faux Positif: 1 (0.00%)
- Faux Négatif: 9 (0.01%)
- Vrai Positif: 62107 (99.99%)

Importance des variables :
Ces observations nous ont contraint à identifier davantage les variables qui déterminent sans équivoque la température. Voici le graphique que nous obtenons:



Nous pouvons ainsi voir que la densité et la salinité sont deux variables déterminantes dans la classification de la température, grâce aux résultats ci-contre :

- Densité: coefficient négatif très important (≈ -300)
- Salinité: coefficient positif (≈ 100)
- Profondeur et Oxygène: coefficients proches de zéro

La courbe ROC (Image 4) et la distribution des probabilités prédites (Image 3) montrent une séparation presque parfaite des classes, avec une AUC de 1.00.

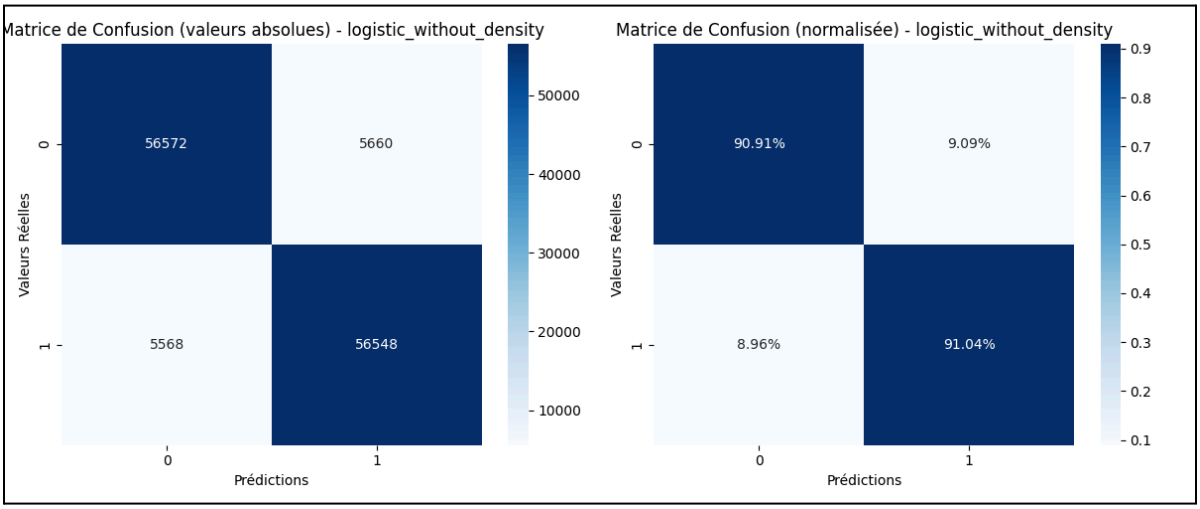
3.1.2 Modèle sans densité

Une fois avoir identifié les variables gênantes dans l'entraînement du modèle, nous nous sommes intéressés à la nature des variables, afin d'identifier quelle variable rend impertinente la démarche. Grâce à nos recherches nous avons donc remarqué que la densité était le résultat de la salinité multipliée par la température. Nous l'avons donc retiré de ce test.
Le deuxième entraînement concernait ainsi le modèle de régression logistique en utilisant uniquement les variables Salinité, Profondeur et Oxygène.

Voici les résultats de la **Performance** du modèle au bout de 5000 itérations :

- Accuracy: 0.9097
- Precision: 0.9090
- Recall: 0.9104
- F1 Score: 0.9097
- AUC-ROC: 0.9695

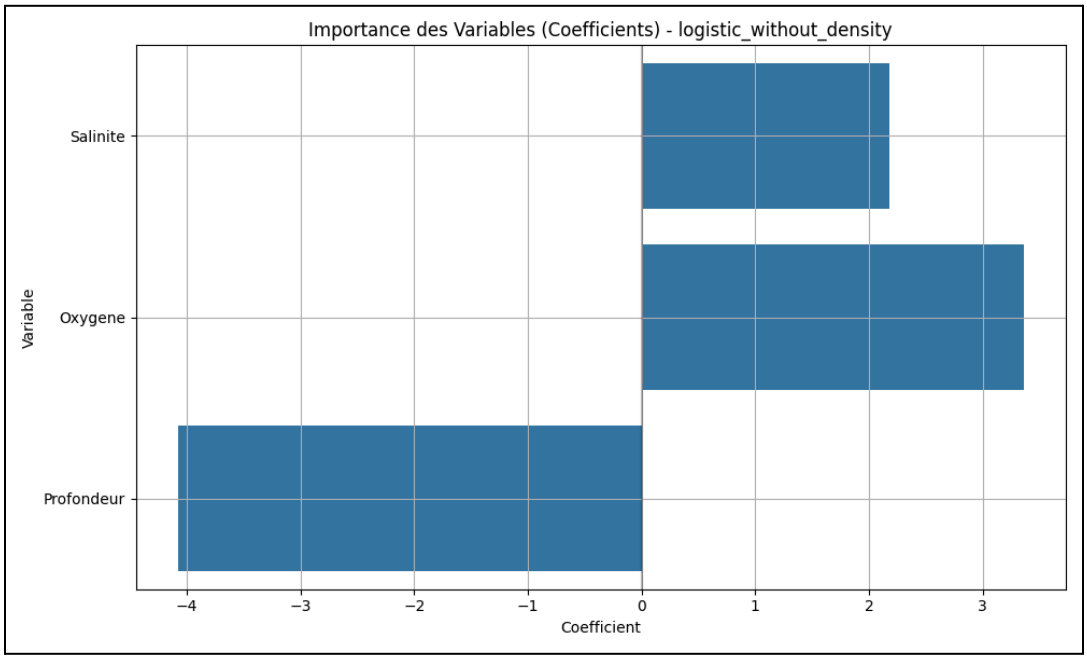
Matrice de confusion :



A partir de cette matrice, nous pouvons faire les observations suivantes:

- Vrai Négatif: 56572 (90.91%)
- Faux Positif: 5660 (9.09%)
- Faux Négatif: 5568 (8.96%)
- Vrai Positif: 56548 (91.04%)

Importance des variables



Voici les coefficients obtenus après la régression logistique :

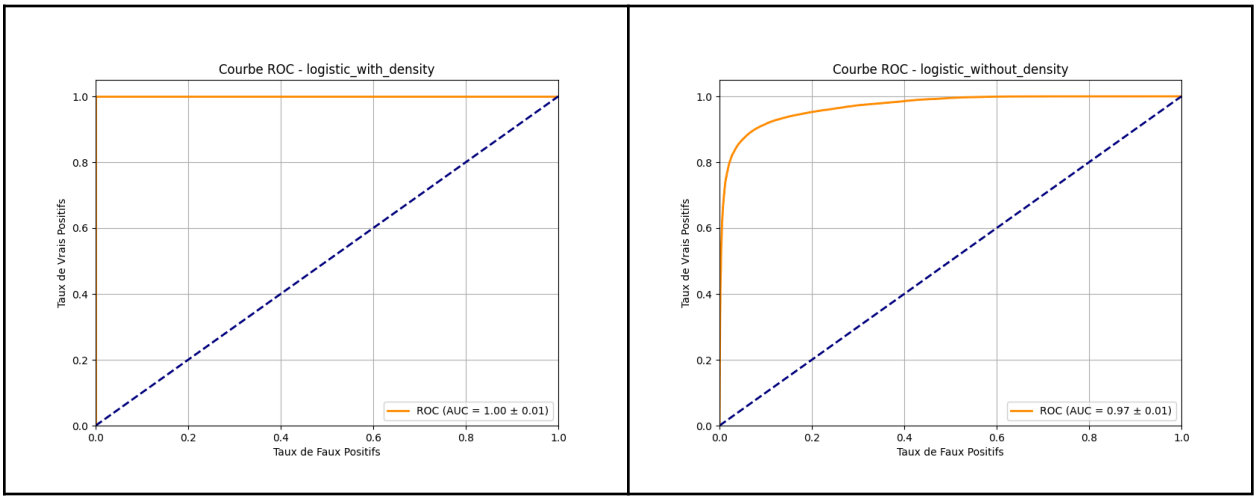
- Oxygène: coefficient positif important (≈ 3.0)
- Salinité: coefficient positif modéré (≈ 2.0)
- Profondeur: coefficient négatif significatif (≈ -4)

3.1.3 Comparaison des modèles de classification

La différence spectaculaire de performance entre les deux modèles (91% vs. 100% d'accuracy) nous a surpris, cela nous a poussé à nous renseigner sur les méthodes de calcul de cette variable, après des recherches plus approfondies cela s'explique par la relation circulaire créée par l'inclusion de la densité, qui est directement calculée à partir de la température et de la salinité.

3.1.4 Mesure des performances

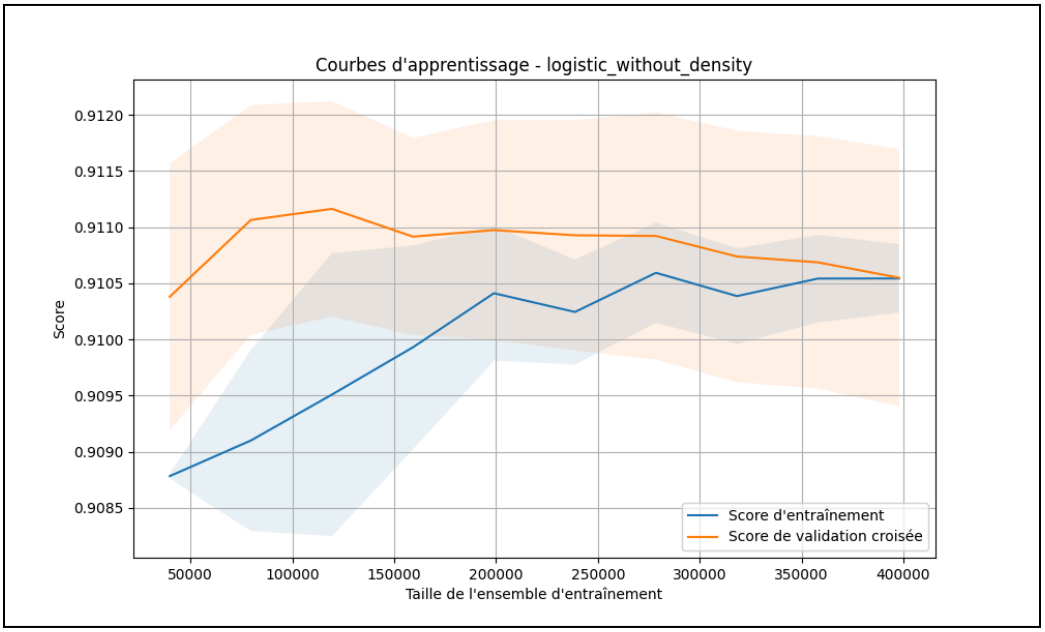
Nous avons utilisé la courbe ROC pour les deux jeux de variables utilisés :



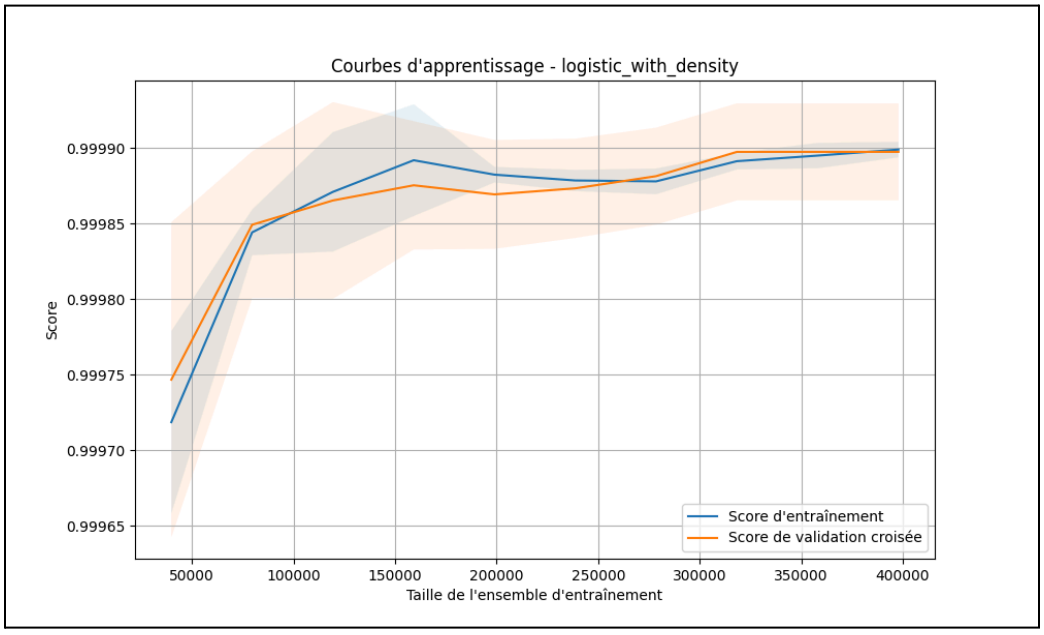
La première courbe nous confirme bien que le jeu de variables utilisé n'était pas le bon dans l'entraînement du modèle.
La deuxième courbe peut nous aider, à partir d'un seuil de confiance par exemple a = 0.95, à identifier le jeu de données suffisant pour entraîner convenablement le modèle.

Dans le but d'identifier les performances du modèle en fonction du nombre de données d'entraînement, nous avons construit les courbes d'apprentissage des modèles suivantes :

- Modèle sans densité : légère amélioration avec plus de données, pas de sur-apprentissage car la courbe d'entraînement ne dépasse pas la courbe de validation.



- Modèle avec densité : performance presque parfaite dès le départ :



3.2 Régression Continue

La classification ne prend pas en compte toute la complexité de la situation, nous considérons donc pertinent d'utiliser la régression continue dans cette situation pour affiner les résultats du modèle.

3.2.1 Régression linéaire

Nous avons entraîné un modèle de régression linéaire pour prédire la température numérique exacte.

Performances:

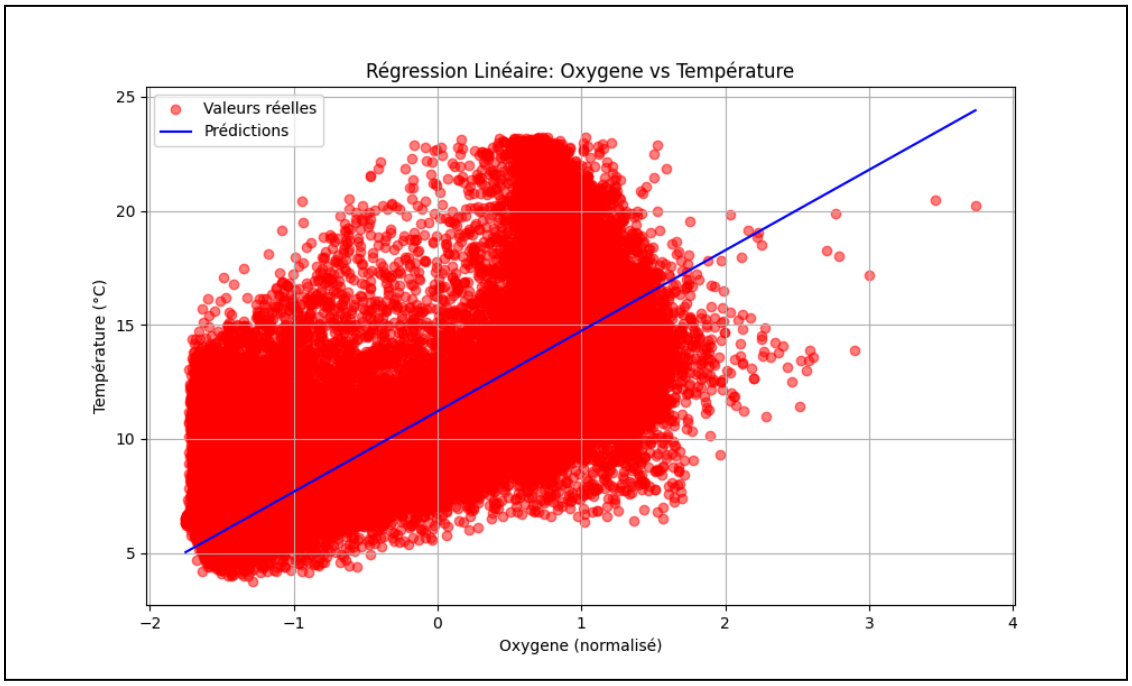
- R^2 Score: 0.7624
- RMSE: 1.8778°C

Coefficients du modèle:

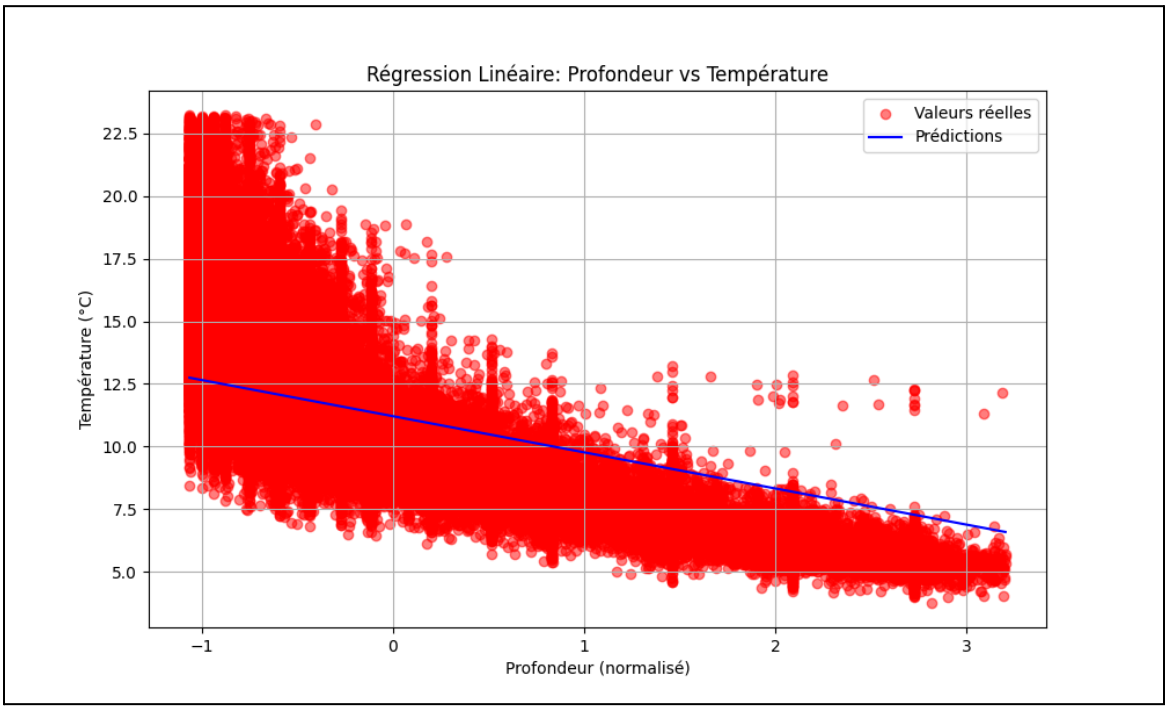
- Salinite: 1.9905
- Profondeur: -1.4444
- Oxygène: 3.5272
- Constante: 11.2106

La visualisation de la relation entre chaque variable et la température (Images 2, 4 et 9) montre que:

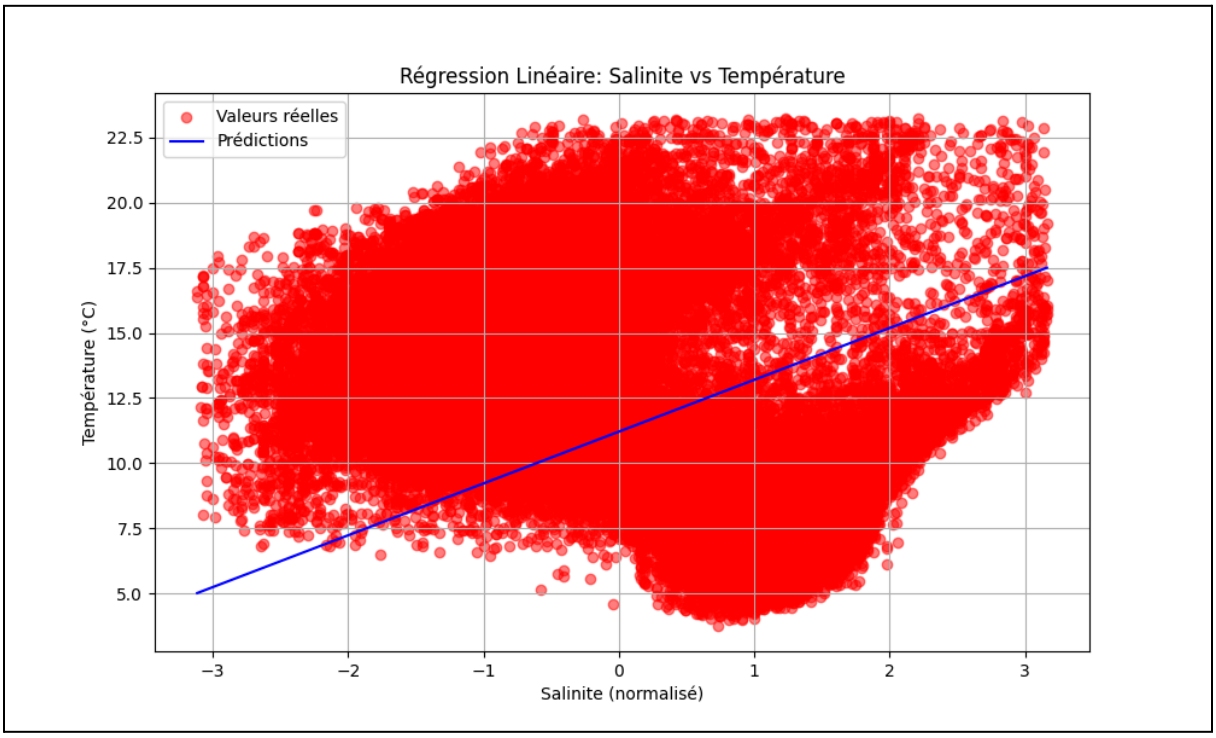
- La concentration en oxygène présente une relation positive avec la température, mais comporte une dispersion importante. Cette complexité démontre les limites d'un modèle linéaire.



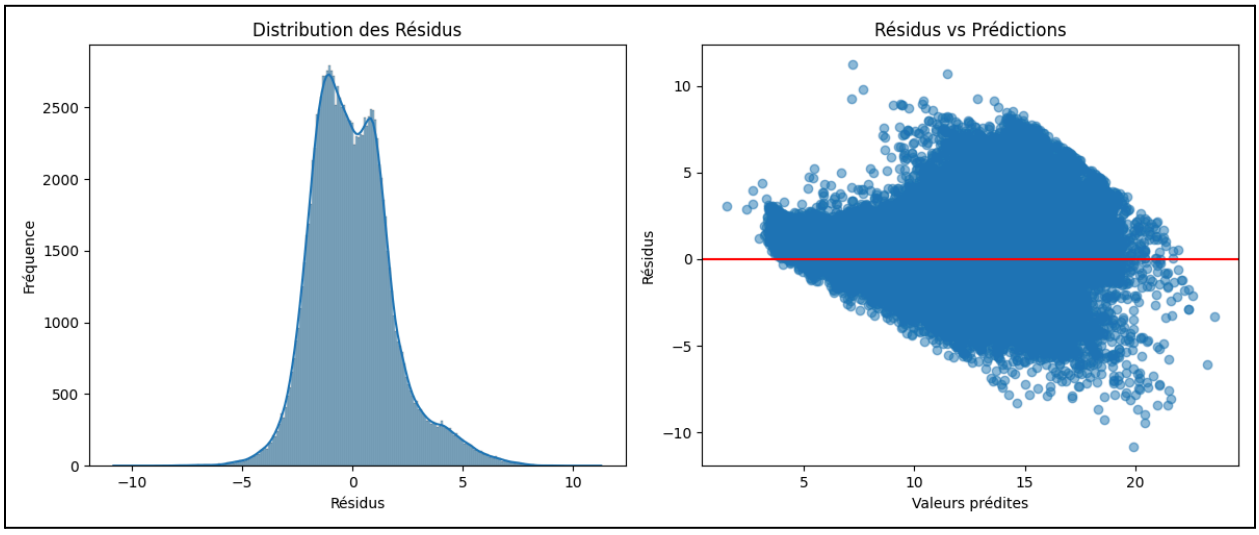
- La profondeur montre une relation négative linéaire avec la température.



- La salinité présente une relation positive avec la température, mais non-linéaire.



Etude de la dispersion des valeurs par rapport aux prédictions:



L'analyse des résidus de la régression linéaire montre une distribution relativement normale, mais avec une tendance visible dans les résidus par rapport aux valeurs prédites, ce qui suggère la présence de relations non-linéaires, et visiblement non capturées par le modèle. Cette distribution presque normale mais pas parfaite est un premier indice que le modèle n'est pas optimal.

3.2.2 Régression polynomiale

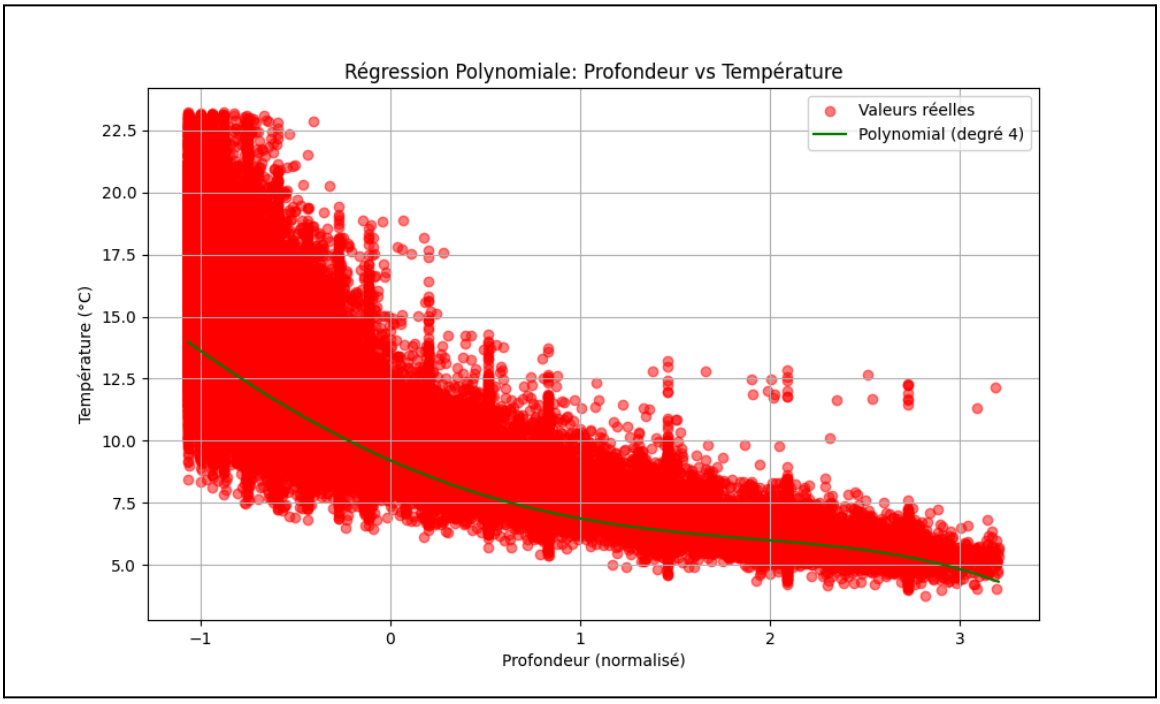
Nous avons donc testé un modèle de régression polynomiale de degré 4 pour obtenir un modèle dont les résultats correspondent davantage au réel, le modèle de régression linéaire étant insuffisant et trop éloigné du réel.

Performances:

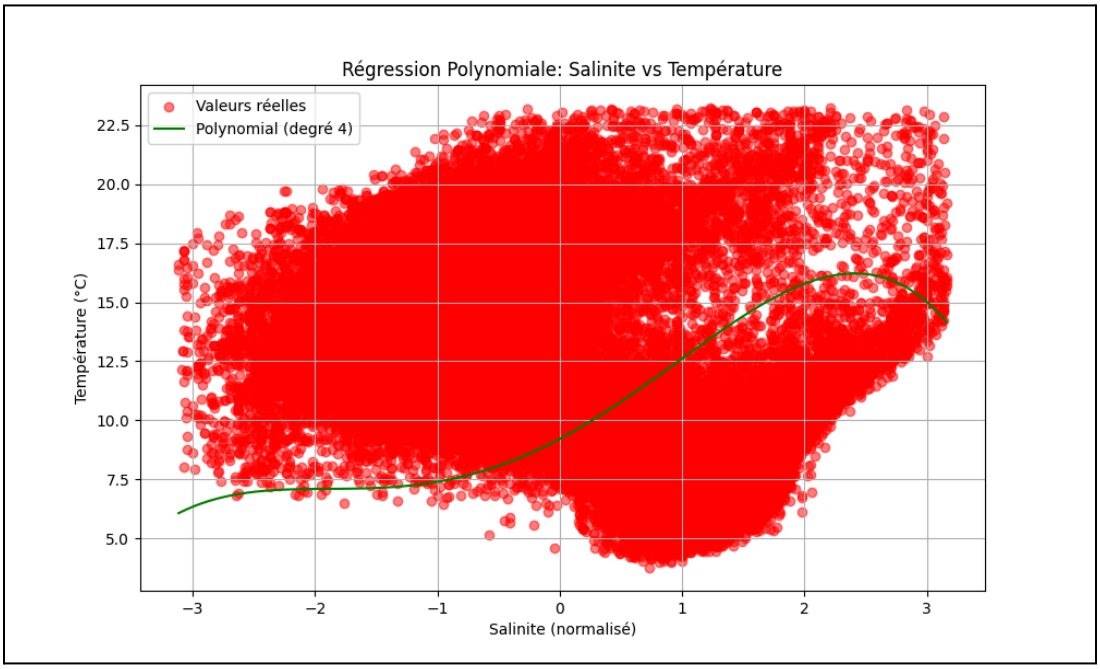
- R^2 Score: 0.8740
- RMSE: 1.3676°C

Les visualisations des régressions polynomiales montrent une meilleure capacité à capturer les relations non-linéaires:

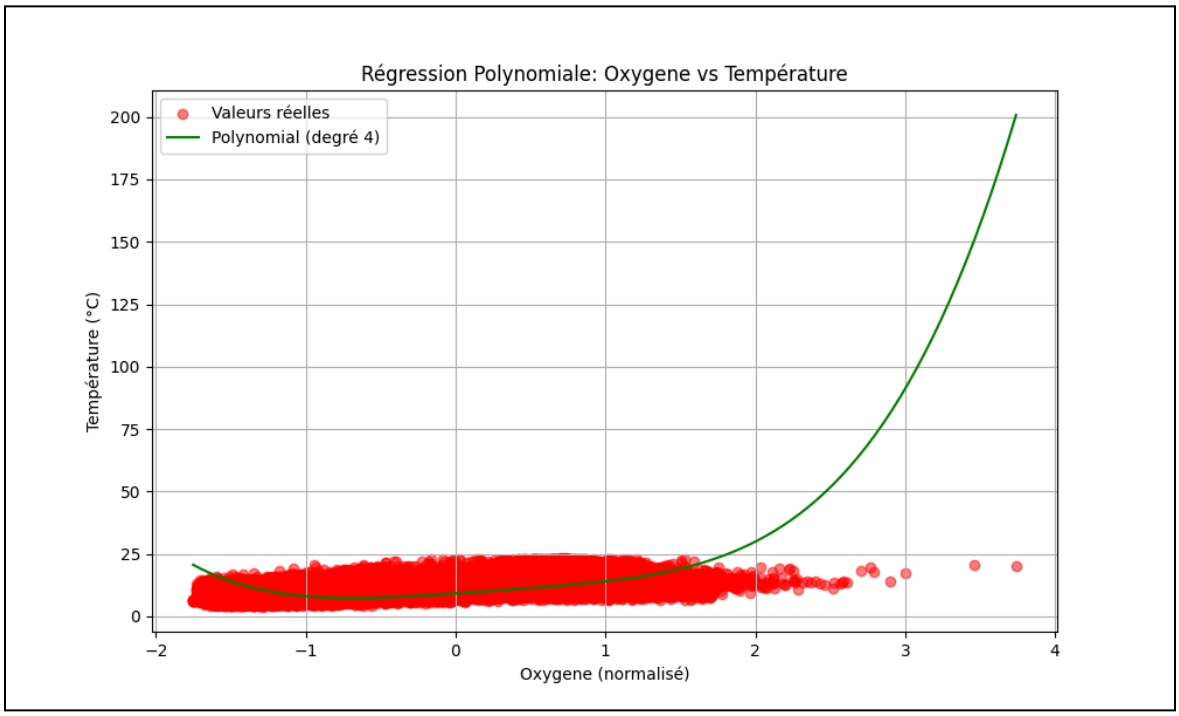
- La relation profondeur-température suit davantage les valeurs réelles par une courbe:



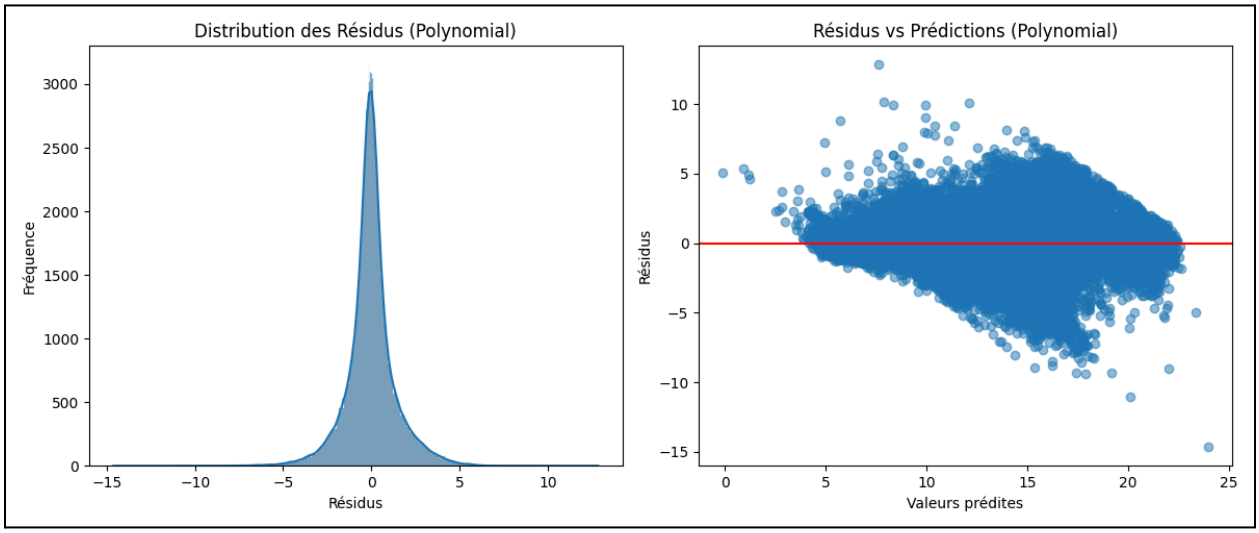
- La salinité présente une relation complexe en forme de "S" avec la température:



- La relation oxygène-température montre un comportement exponentiel pour les valeurs extrêmes

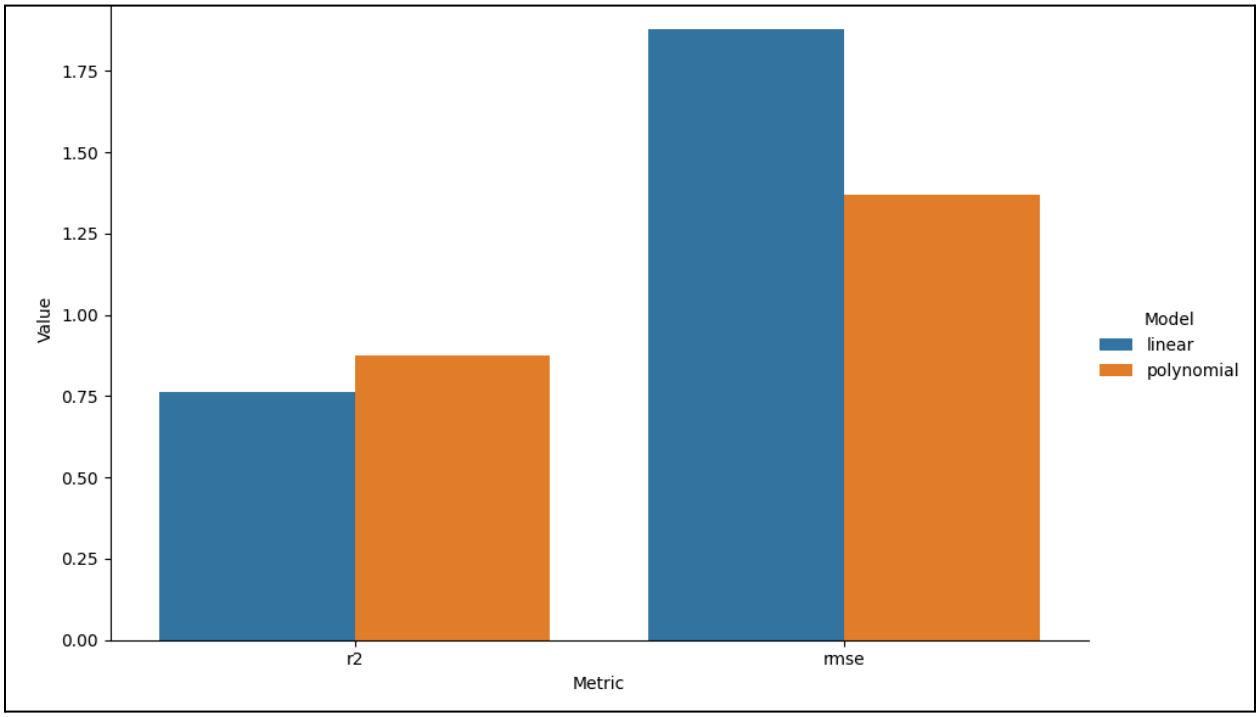


Etude de la dispersion des valeurs par rapport aux prédictions:



L'analyse des résidus du modèle polynomial montre une distribution plus centrée et moins de structure dans le graphique résidus et prédictions, indiquant une meilleure capture des relations sous-jacentes.

3.2.3 Comparaison des modèles de régression



Graphique comparatif des modèles de régression.

La régression polynomiale surpasse significativement la régression linéaire, avec une amélioration de l'ordre de 11% pour le R^2 et une réduction de 27% de l'erreur (RMSE), comme le montre clairement le graphique comparatif. Ces résultats indiquent la présence de relations non linéaires importantes entre les variables explicatives et la température.

4. Interprétation des Résultats

4.1 Classification: Impact de la Densité

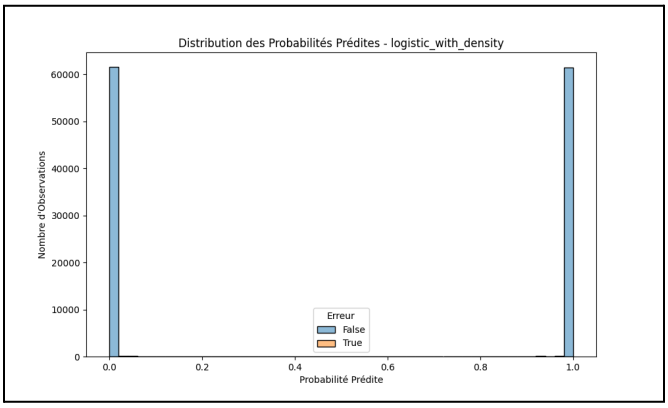
La comparaison des deux modèles de classification révèle un phénomène crucial:

- **Avec densité**: Accuracy = 0.9999
- **Sans densité**: Accuracy = 0.9097

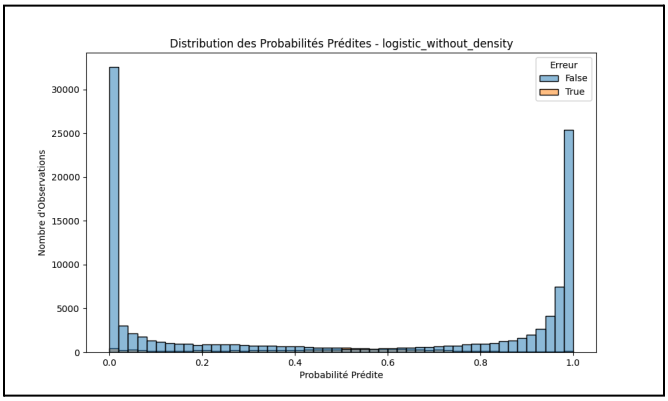
Cette différence spectaculaire s'explique par une relation circulaire: **la densité est calculée directement à partir de la température et la salinité** via l'équation d'état de l'eau de mer. En incluant la densité comme variable explicative, nous introduisons implicitement la température elle-même dans le modèle.

Les visualisations confirment cette analyse:

- La distribution des probabilités prédites avec densité montre une séparation parfaite (0 ou 1)



- L'importance des variables montre la dominance écrasante de la densité



4.2 Analyse des Variables pour la Classification sans Densité

L'analyse des variables dans le modèle sans densité permet une interprétation physique cohérente:

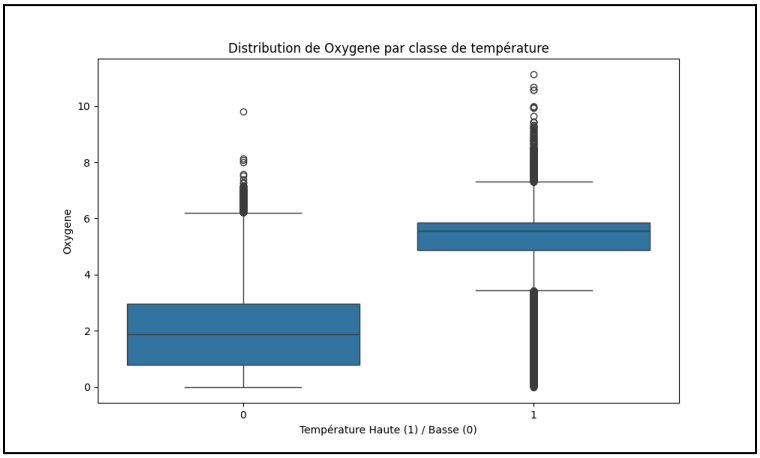
1. **Oxygène (coefficient positif)**: Les eaux plus riches en oxygène sont généralement associées à des températures plus élevées près de la surface, où l'échange avec l'atmosphère et la photosynthèse sont plus actifs.

2. **Profondeur (coefficient négatif)**: La température diminue avec la profondeur, ce qui explique le coefficient négatif significatif.

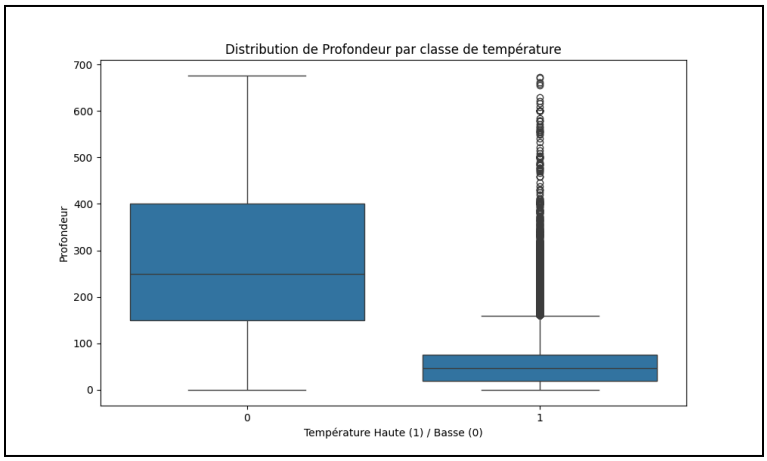
3. **Salinité (coefficient positif)**: Dans ce dataset, la salinité montre une relation positive avec la température, bien que cette relation puisse varier selon les régions océaniques.

Les boxplots des distributions de variables par classe suivants confirment ces relations:

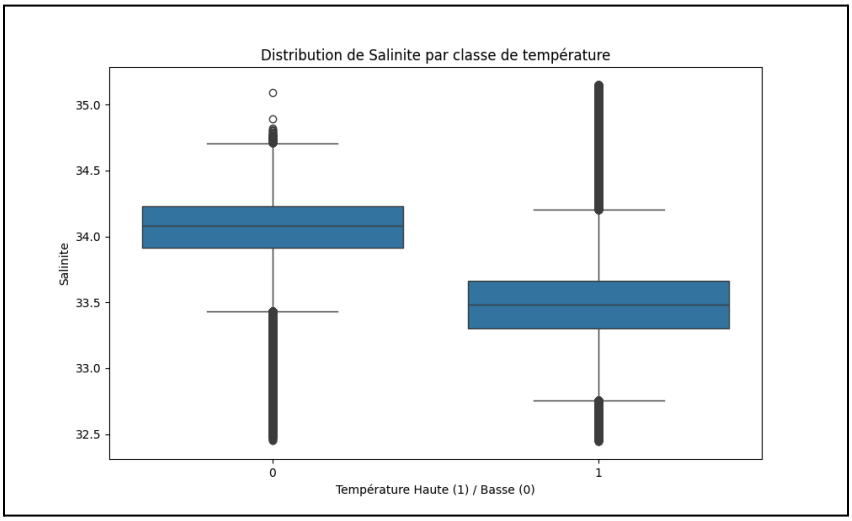
- Oxygène plus élevé dans les eaux chaudes:



- Profondeur plus faible pour les températures élevées:



- Salinité légèrement plus faible dans les eaux chaudes:



4.3 Régression: Performances et Interprétabilité

La régression linéaire atteint un R^2 de 0.7624, tandis que le modèle polynomial atteint 0.8740. Cette amélioration avec la régression polynomiale indique la présence de relations non linéaires entre les variables océanographiques, comme le confirme visuellement la comparaison des performances.

Les coefficients du modèle linéaire permettent d'interpréter l'impact de chaque variable:

- Oxygène (3.53): forte relation positive avec la température, coefficient le plus important
- Salinité (1.99): relation positive modérée avec la température
- Profondeur (-1.44): effet négatif sur la température (plus profond = plus froid)

L'analyse visuelle des relations variables-température confirme ces tendances mais révèle également leur nature non-linéaire:

- La profondeur montre une relation décroissante linéaire imparfaite avec la température
- La salinité présente une relation complexe en forme de "S" avec la température, difficile à capturer avec un modèle linéaire
- L'oxygène montre une tendance clairement exponentielle aux valeurs élevées

L'analyse des résidus confirme la supériorité du modèle polynomial, avec une distribution plus centrée et moins de structure dans les résidus.

Ces observations sont cohérentes avec les principes océanographiques connus et avec les corrélations observées dans les données.

4.4 Limites de l'Étude

Plusieurs limitations importantes doivent être considérées:

1. **Relation circulaire avec la densité**: L'inclusion de cette variable crée un modèle artificiellement parfait

2. **Variables manquantes**: D'autres facteurs influencent la température océanique mais ne sont pas inclus dans notre modèle:
    - Saisonnalité et variations temporelles
    - Courants marins

3. **Absence d'exploration d'autres algorithmes**: Nous nous sommes limités à la régression logistique et linéaire/polynomiale, sans explorer d'autres méthodes comme les forêts aléatoires ou les SVM.

5. Conclusion

Cette étude a démontré l'efficacité des techniques de machine learning pour analyser les relations complexes entre variables océanographiques et prédire la température de l'eau.

Points clés:

1. **Classification binaire**: Même sans intégrer la densité, la prédiction des températures basses/hautes atteint environ 91 % de précision, ce qui est tout à fait respectable et réaliste.

2. **Régression continue**: La prédiction précise de la température est possible avec $R^2 > 0.87$ en utilisant un modèle polynomial. Autrement dit, plus de 87 % de la variation observée dans les températures est expliquée par le modèle, ce qui montre une bonne capacité prédictive du modèle.

3. **Impact de la densité**: La relation circulaire créée par l'inclusion de la densité nous a appris l'importance de comprendre les relations physiques entre les variables pour éviter de chercher des erreurs dans notre code durant des heures.

4. **Relations non linéaires**: Les performances supérieures du modèle polynomial confirment la nature non linéaire des interactions entre variables océanographiques.

5. **Importance des variables**: L'oxygène et la profondeur sont les prédicteurs les plus importants de la température dans les modèles sans densité.