Rapport méthodologique

Prédiction des précipitations corrigées

Hackathon Zindi – NK STAT CONSULTING

1er août 2025

Table des matières

1. Contexte et Problématique

Ce projet de hackathon Zindi vise à prédire les précipitations corrigées à partir de variables météorologiques dans la ville de Brazzaville. La prédiction précise des précipitations est cruciale pour la gestion des ressources en eau, l'agriculture et la prévention des risques climatiques. L'approche adoptée combine l'expertise météorologique avec des techniques avancées d'apprentissage automatique pour créer un modèle robuste et scientifiquement fondé.

La méthodologie suit un pipeline structuré : données brutes → ingénierie des caractéristiques → analyse exploratoire → clustering → modélisation → validation. Cette approche holistique permet de capturer à la fois les relations physiques connues en météorologie et les patterns complexes révélés par les données.

2. Ingénierie des caractéristiques : Une approche scientifiquement fondée

Le dataset initial contient les variables météorologiques classiques comme la vitesse du vent (`WS2M`), la température (`T2M`), l'humidité relative (`RH2M`), la pression de surface (`PS`) et l'humidité spécifique (`QV2M`). Cependant, l'innovation réside dans la création de variables dérivées basées sur la physique atmosphérique.

Les variables thermodynamiques créées incluent le ratio de saturation (`HUMIDITY_SATURATION = QV2M / (RH2M/100)`), l'écart au point de rosée (`DEW_POINT_SPREAD = T2M - T2MDEW`) et la différence de température du bulbe humide. Ces indicateurs captent respectivement la capacité d'absorption d'eau de l'air, le potentiel de condensation et la capacité évaporative – tous essentiels pour comprendre les processus de formation des précipitations.

L'encodage temporel utilise des transformations sinusoïdales et cosinusoïdales pour les mois et jours, par exemple :

$$\texttt{MONTH\_SIN} = \sin\left(\frac{2\pi \times \texttt{MO}}{12}\right)$$

Cette approche cyclique évite la discontinuité artificielle des variables temporelles classiques et respecte la nature cyclique des phénomènes météorologiques. Les saisons sont également codées selon les spécificités climatiques du Congo-Brazzaville : petite saison sèche (janvier-février), grande saison sèche (juin-septembre), grande saison des pluies (octobre-décembre) et petite saison des pluies (mars-mai).

3. Analyse exploratoire : comprendre les données en profondeur

L'analyse exploratoire suit une approche systématique en trois phases. D'abord, l'analyse univariée examine chaque variable individuellement avec des histogrammes pour les variables quantitatives et des graphiques en barres pour les variables qualitatives. Cette étape révèle les distributions, identifie les valeurs aberrantes et caractérise la variabilité de chaque variable.

L'analyse de la variable cible (précipitations) montre une distribution avec des outliers significatifs. Bien que la méthode IQR (Interquartile Range) soit utilisée pour identifier statistique-

ment ces valeurs extrêmes, la décision de filtrage retient spécifiquement les observations dont la variable cible est inférieure à 50mm (`train = train.query('Target<=50')`). Ce seuil de 50mm permet d'éliminer les événements météorologiques exceptionnels qui pourraient être dus à des erreurs de mesure ou à des phénomènes non représentatifs du climat habituel, tout en conservant suffisamment de données pour l'entraînement du modèle.

L'analyse saisonnière discriminante utilise des box plots et des tests de Kruskal-Wallis pour identifier quelles variables diffèrent significativement selon les saisons. Cette analyse guide la sélection des caractéristiques les plus pertinentes pour la modélisation en révélant les variables qui capturent le mieux la variabilité saisonnière des précipitations.

4. Analyse multivariée et identification de régimes météorologiques

L'Analyse en Composantes Principales (ACP) révèle la structure sous-jacente des variables météorologiques. Les deux premières composantes capturent 68 % de la variance totale, avec la première composante (46.5 %) représentant principalement les relations température-saisonnalité et la seconde (21 %) les interactions vent-pression-humidité. Le cercle des corrélations montre des groupes cohérents de variables évoluant ensemble.

Le clustering K-means appliqué sur l'espace ACP identifie quatre régimes météorologiques distincts. Cette approche non supervisée révèle des patterns naturels dans les données correspondant probablement à différentes situations synoptiques : conditions anticycloniques versus dépressionnaires, masses d'air sèches versus humides. L'utilisation de la méthode du coude optimise le nombre de clusters, et l'application des mêmes centres sur les données de test assure la cohérence.

5. Stratégie de modélisation : Comparaison multi-algorithmes

La modélisation compare cinq algorithmes différents : régression linéaire (baseline), Ridge (régularisation L2), Random Forest (ensemble non-linéaire), Gradient Boosting et XGBoost. Cette approche comparative permet d'identifier l'algorithme le mieux adapté aux spécificités des données météorologiques.

Le preprocessing utilise un `ColumnTransformer` avec `RobustScaler` pour les variables numériques (résistant aux outliers) et `OneHotEncoder` pour les variables catégorielles (évitant la multicolinéarité). Cette approche standardisée assure que tous les modèles opèrent sur des données comparables.

L'optimisation par `GridSearch` avec validation croisée 5-fold affine les hyperparamètres de chaque modèle. Pour Gradient Boosting par exemple, les paramètres optimisés incluent le nombre d'arbres (50-200), le taux d'apprentissage (0.01-0.2) et la profondeur maximale (3-7). Cette optimisation systématique maximise les performances tout en évitant le sur-apprentissage.

6. Résultats et performance : Le Gradient Boosting Comme Solution Optimale

L'évaluation utilise plusieurs métriques complémentaires : R^2 (variance expliquée), RMSE (erreur quadratique moyenne) et MAE (erreur absolue moyenne). La comparaison train/test permet de détecter le sur-apprentissage, crucial pour assurer la généralisation du modèle.

Le Gradient Boosting optimisé émerge comme la solution la plus performante, montrant le meilleur équilibre entre précision et généralisation. L'analyse de l'importance des caractéristiques du modèle Gradient Boosting optimal doit être interprétée à partir des résultats réels affichés par `feature_importances['GradientBoosting']`. Cette analyse révélera quelles variables contribuent le plus aux prédictions : variables météorologiques physiques (`WS2M`, `T2M`, `RH2M`, `PS`, `QV2M`), variables temporelles cycliques (`MONTH_SIN`, `MONTH_COS`), ou variables contextuelles (classe des régimes météorologiques, `SEASON`).

La validation temporelle montre que les prédictions suivent une évolution saisonnière cohérente, confirmant que le modèle a appris les patterns climatiques sous-jacents plutôt que de simples corrélations spurieuses.

Note méthodologique importante : L'interprétation précise de la contribution de chaque variable nécessite l'examen des résultats réels de `feature_importances['GradientBoosting']` qui classe les variables par ordre d'importance décroissante. Cette analyse déterminera si ce sont les variables physiques, temporelles ou contextuelles qui dominent la prédiction.

7. Forces de la méthodologie et perspectives d'amélioration

Cette approche se distingue par sa fondation scientifique solide, combinant expertise météorologique et techniques modernes d'apprentissage automatique. L'ingénierie des caractéristiques respecte les principes physiques (équation de Clausius-Clapeyron, processus adiabatiques), tandis que l'analyse multivariée révèle des structures cachées dans les données.

Les principales forces incluent : la robustesse du pipeline de traitement, la comparaison systématique multi-algorithmes, l'optimisation rigoureuse des hyperparamètres et la validation multiple. Cette méthodologie est reproductible, scalable et scientifiquement fondée.

Les améliorations possibles incluent l'ajout de variables d'interaction (`T2M * RH2M`), l'utilisation de variables retardées pour capturer les tendances, l'implémentation d'ensembles hybrides (stacking) et l'adoption de techniques d'interprétabilité avancées (SHAP values) pour une analyse plus fine de la contribution de chaque variable. Une validation temporelle stricte (train sur années passées, test sur année récente) renforcerait également la robustesse opérationnelle du modèle.

8. Justification de la non-utilisation des modèles séquentiels (LSTM)

Bien que les réseaux de neurones récurrents comme les LSTM (Long Short-Term Memory) soient une référence pour l'analyse de séries temporelles, leur non-utilisation dans ce projet est un choix méthodologique délibéré reposant sur plusieurs arguments.

1. **Transformation du problème en données tabulaires :** Notre approche repose sur une ingénierie des caractéristiques poussée qui transforme le problème séquentiel en une tâche de régression supervisée. En encodant explicitement la temporalité (transformations sinusoïdales pour les mois et jours, variables saisonnières), l'information séquentielle est déjà capturée, rendant l'apprentissage implicite par un LSTM redondant.
2. **Performance et interprétabilité des modèles ensemblistes :** Sur des données structurées et hétérogènes comme les nôtres, les algorithmes ensemblistes (Gradient Boosting, Random Forest) sont souvent plus performants et efficaces. Surtout, ils offrent une interprétabilité directe via l'importance des caractéristiques (`feature_importances_`), un élément central de notre démarche visant à fonder scientifiquement le modèle. L'analyse des mécanismes internes d'un LSTM est notoirement plus complexe.
3. **Efficacité pragmatique :** Dans le contexte d'un développement rapide comme un hackathon, les modèles ensemblistes présentent un meilleur compromis entre le temps d'entraînement, le coût de calcul et la performance prédictive par rapport aux réseaux de neurones profonds qui exigent une optimisation plus lourde.

Ainsi, le choix du Gradient Boosting s'est avéré plus aligné avec les objectifs de performance, d'interprétabilité et d'efficacité du projet.

9. Conclusion

Cette méthodologie représente une approche exemplaire de la prédiction météorologique moderne, alliant rigueur scientifique et innovation technique. Le succès du Gradient Boosting optimisé démontre l'efficacité de cette approche intégrée pour capturer les relations complexes entre variables météorologiques et précipitations, ouvrant la voie à des applications opérationnelles robustes.