République du Cameroun

Paix-Travail-Patrie

MINSUP

Université de Douala

Faculté des Sciences

Republic of Cameroon

Peace-Work-Fatherland

MINSUP

University of Douala

Faculty Of Science

TPE INF 365 – GROUPE 24

Étudiants :	Matricules :
• NITOPOP JEATSA GUILLAUME MELVIN (CHEF)	20S43003
• DEMANOU KEMKENG DILAN	
• NOSSI YIMGO LYNDSEY SULIVANE	23S87713
• TCHIEUTCHOUA FOTEPING ASHLEY MEGANE	23S87863
• WANGA POUYA KAVEN SAMIRA	23S88070

Thème :

Prévision des crues – Etat d'avancement

Examinateur :

Dr Justin MOSKOLAI

Sommaire

Introduction

Dans le monde, les inondations sont devenues un problème récurrent qui nous causent des pertes humaines et matérielles dans nos villes. Mais nous essayons de les prévenir à Douala. C'est pourquoi dans l'œuvre de ce travail pratique (TP), nous allons présenter des modèles de machine learning[(1)] qui peuvent nous permettre de calculer les probabilités d'inondation dans une région donnée.

I. Objectifs :

Ce TP vise à développer un modèle de machine learning[(2)] capable de prédire les crues à partir de données météorologiques et hydrologiques. L'objectif est de :

- Comprendre les facteurs influençant les crues.
- Appliquer les étapes classiques d'un projet de data science[(3)] : préparation des données, modélisation, évaluation et interprétation.
- Comparer plusieurs modèles pour identifier celui offrant les meilleures performances.

Ce projet a donc un problème de régression[(4)] car nous devons pouvoir ressortir à partir des données d'entrée les risques d'inondation.

II. Contexte :

Les crues représentent un risque majeur pour les populations et les infrastructures. Pouvoir les anticiper permettrait de mieux gérer les évacuations, les ressources et les alertes. Ce TP s'inscrit dans une démarche de prévention des catastrophes naturelles grâce à l'intelligence artificielle[(5)].

III. Étapes du projet :

a. Préparation des données :

Le jeu de données comprend 21 variables numériques telles que :

- Intensité de la mousson
- Topographie et drainage
- Gestion des rivières
- Déforestation
- Urbanisation
- Changement climatique
- Qualité des barrages
- Envasement
- Pratiques agricoles
- Empiètements
- Préparation inefficace aux catastrophes
- Systèmes de drainage
- Vulnérabilité côtière
- Glissements de terrain
- Bassins versants
- Dégradation des infrastructures
- Indice de population
- Perte des zones humides
- Planification inadéquate
- Facteurs politiques
- Probabilité d'inondation

Expliquons l'effet de chaque variable :

Variable (Français)	Impact sur les crues
Intensité de la mousson	Des pluies intenses sur une courte période augmentent le risque de crue.
Topographie et drainage	Les terrains plats ou mal drainés favorisent l'accumulation d'eau.

Gestion des rivières	Une mauvaise gestion (digues, curage) peut provoquer des débordements.
Déforestation	Réduit l'absorption de l'eau par les sols, augmente le ruissellement.
Urbanisation	Les surfaces imperméables (béton, asphalte) empêchent l'infiltration.
Changement climatique	Accroît la fréquence et l'intensité des événements pluvieux extrêmes.
Qualité des barrages	Des barrages défectueux peuvent céder ou mal réguler les crues.
Envasement	Réduit la capacité des rivières à contenir l'eau, favorisant les crues.
Pratiques agricoles	Peuvent compacter le sol ou favoriser l'érosion, réduisant l'absorption.
Empiètements	L'occupation des zones inondables bloque l'écoulement naturel.
Préparation inefficace aux catastrophes	Aggrave les impacts des crues par manque d'alerte ou d'organisation.
Systèmes de drainage	Des canalisations obstruées ou mal conçues ne permettent pas l'évacuation rapide.
Vulnérabilité côtière	Expose aux crues liées aux tempêtes et à la montée des eaux.
Glissements de terrain	Peuvent bloquer les rivières et créer des crues soudaines.
Bassins versants	Leur état influence la répartition et la rétention de l'eau.
Dégradation des infrastructures	Routes, ponts ou canalisations vétustes peuvent aggraver les inondations.
Indice de population	Une population dense augmente les risques humains et la pression sur les systèmes.
Perte des zones humides	Diminue la capacité naturelle d'absorption et de régulation de

	l'eau.
Planification inadéquate	L'aménagement du territoire sans prise en compte du risque hydrologique crée des zones vulnérables.
Facteurs politiques	Influencent les décisions sur la gestion des ressources et des risques.
Probabilité d'inondation	Résultat final du modèle, basé sur l'ensemble des variables ci-dessus.

Sur ces données, quelques traitements seront effectués tel que :

- Nettoyage : suppression des valeurs manquantes, des doublons et des outliers[6].
- Exploration :
  - Corrélation[7] forte entre précipitations et crues.
  - Les crues surviennent souvent après plusieurs jours de fortes pluies.
- Normalisation[8] : standardisation[9] des variables continues pour les modèles sensibles à l'échelle.

b. Implémentation[10] des modèles :

Trois modèles ont été testés :

- Régression linéaire[11] : simple et interprétable.
- Forêt aléatoire[12] : robuste et efficace pour les données tabulaires.
- Réseau de neurones[13] séquentiel[14] : bon pour les problèmes de régression.

Les données ont été divisées en 80 % pour l'entraînement et 20 % pour le test.

c. Évaluation des performances :

Modèle	MSE[15]	RMSE[16]	MAE[17]	R^2-score[18]
Régression logistique	0.000014308	0.003782	0.002189	0.9942
Forêt aléatoire	0.0006683	0.0258	0.0204	0.7316
Réseau de neurones séquentiel	0.0000136	0.0036907	0.0023931	0.9945

Le réseau de neurones séquentiel a obtenu les meilleurs résultats globaux.

d. Analyse des résultats :

- Les précipitations et le débit sont les variables les plus importantes.
- Le modèle de forêt aléatoire n'est pas assez performant à cause de variables peu informatives qui sont peu corrélées à la cible.
- Le réseau de neurones séquentiel a la meilleure performance car il peut mieux capter les interactions complexes entre les variables.
- La régression linéaire est simple mais moins précise par rapport aux réseaux de neurones séquentiels.

IV. Expérimentations supplémentaires :

- Validation croisée[19] : 5-fold CV[20] utilisée pour éviter le surapprentissage[21].
- Tuning[22] d'hyperparamètres[23] :[24] GridSearchCV[25] appliqué à la forêt aléatoire.
- Test d'un modèle XGBoost[26] : performances similaires à la forêt aléatoire.

V. Réflexions :

- Pourquoi le réseau de neurones séquentiel fonctionne bien ? Elle gère bien les interactions complexes entre plusieurs variables.
- Améliorations possibles :
    - Intégrer des données temporelles (séries chronologiques).
    - Tester des réseaux de neurones récurrents[27] (RNN, LSTM[28]).

VI. Conclusion :

Ce TP m'a permis de comprendre les étapes clés d'un projet de machine learning appliqué à un problème environnemental réel. Le réseau de neurones séquentiel s'est révélé être le modèle le plus performant pour la prédiction des crues. Ce type d'approche pourrait être intégré dans des systèmes d'alerte précoce pour sauver des vies.

VII. Annexes :

1. Glossaire des termes techniques :

1. Machine Learning : Branche de l'intelligence artificielle qui permet à des systèmes informatiques d'apprendre à partir de données sans être explicitement programmés pour chaque tâche.
2. Modèle de Machine Learning : Structure mathématique ou algorithmique qui apprend à partir de données pour faire des prédictions ou prendre des décisions sans être explicitement programmée pour chaque tâche.
3. Data Science : Domaine interdisciplinaire qui combine statistiques, informatique et connaissance métier pour extraire des informations utiles à partir de données.
4. Régression : Technique utilisée pour prédire une valeur numérique continue à partir de données d'entrée.
5. Intelligence artificielle : Domaine de l'informatique qui vise à créer des systèmes capables de simuler des comportements intelligents, c'est-à-dire

d'apprendre, raisonner, percevoir, comprendre et agir de manière autonome ou assistée.

6. Outliers : 0bservation qui s'écarte fortement des autres données dans un ensemble.
7. Corrélation : Mesure statistique qui indique le degré de relation entre deux variables.
8. Normalisation : Technique utilisée en science des données et en machine learning pour mettre les variables sur une même échelle, afin d'éviter que certaines dominent les autres dans les calculs ou les modèles.
9. Standardisation : Technique de prétraitement des données qui consiste à centrer et réduire les variables, c'est-à-dire à les transformer pour qu'elles aient une moyenne de 0 et un écart-type de 1.
10. Implémentation : Processus par lequel une idée, une méthode ou un algorithme est concrètement réalisée dans un système informatique ou un programme.
11. Régression linéaire : Méthode statistique et de machine learning utilisée pour modéliser la relation entre une variable dépendante (cible) et une ou plusieurs variables indépendantes (features[29]) en supposant que cette relation est linéaire.
12. Forêt aléatoire : Algorithme d'apprentissage supervisé utilisé en classification et en régression. Il repose sur le principe d'ensemble learning, c'est-à-dire qu'il combine plusieurs modèles simples (des arbres de décision[30]) pour produire une prédiction plus robuste et précise.
13. Réseau de neurones : Modèle d'intelligence artificielle inspiré du fonctionnement du cerveau humain.
14. Réseau de neurones séquentiel : Type de modèle d'apprentissage profond dans lequel les couches sont empilées les unes après les autres, de manière linéaire.
15. MSE (Mean Squared Error ou erreur quadratique moyenne) : Mesure de performance utilisée en régression pour quantifier l'écart entre les valeurs prédites par un modèle et les valeurs réelles.
16. RMSE (Root Mean Squared Error ou racine de l'erreur quadratique moyenne) : Mesure d'évaluation utilisée en régression pour quantifier l'écart entre les prédictions d'un modèle et les valeurs réelles, en tenant compte de la magnitude des erreurs.
17. MAE (Mean Absolute Error ou erreur absolue moyenne) : Mesure d'évaluation utilisée en régression pour quantifier l'erreur moyenne entre les prédictions d'un modèle et les valeurs réelles, en prenant la valeur absolue des écarts.
18. R^2 score (ou coefficient de détermination) : Mesure statistique utilisée pour évaluer la qualité d'un modèle de régression.

19. Validation croisée (ou cross-validation) : Technique d'évaluation utilisée en apprentissage automatique pour mesurer la performance d'un modèle de manière plus fiable et éviter le surapprentissage (overfitting).
20. 5-fold CV (ou 5-fold cross-validation ou validation croisée à 5 plis) : Méthode d'évaluation utilisée en machine learning pour tester la robustesse d'un modèle en le confrontant à différentes partitions du jeu de données.
21. Surapprentissage (ou overfitting) : Phénomène en apprentissage automatique où un modèle apprend trop bien les données d'entraînement, au point de mémoriser les détails et le bruit, au lieu de généraliser à de nouvelles données.
22. Tuning (ou ajustement des hyperparamètres) : Processus d'optimisation des paramètres de contrôle d'un modèle pour améliorer ses performances sur un jeu de données donné.
23. Hyperparamètre : Variable de configuration définie avant l'entraînement d'un modèle d'apprentissage automatique.
24. Tuning d'hyperparamètres : Processus d'optimisation des réglages externes d'un modèle d'apprentissage automatique pour améliorer ses performances sur un jeu de données donné.
25. GridSearchCV : Méthode d'optimisation des hyperparamètres utilisée en apprentissage automatique pour trouver la meilleure combinaison de paramètres d'un modèle en testant systématiquement toutes les possibilités dans une grille définie.
26. XGBoost (eXtreme Gradient Boosting) : Bibliothèque open source de machine learning qui implémente une version optimisée de l'algorithme de boosting par gradient[(31)].
27. Réseau de neurones récurrents (RNN, pour Recurrent Neural Network) : Type de réseau de neurones conçu pour traiter des données séquentielles, comme du texte, des séries temporelles ou des signaux audios.
28. LSTM (Long Short-Term Memory) : Type avancé de réseau de neurones récurrent (RNN) conçu pour mémoriser des informations sur de longues séquences.
29. Features (ou caractéristiques, parfois appelées variables explicatives) : Informations ou attributs utilisées par un modèle pour faire des prédictions ou apprendre des relations dans les données.
30. Arbre de décision : Algorithme d'apprentissage supervisé utilisé pour la classification et la régression, qui modélise les décisions sous forme d'une structure arborescente.
31. Algorithme de boosting par gradient : Méthode d'apprentissage supervisé qui combine plusieurs modèles faibles (généralement des arbres de décision) pour créer un modèle puissant, en corrigeant les erreurs de prédiction de manière itérative.

2. Sources de données utilisées :

- **Format** : CSV (floods.csv)
- **Contenu** :
    - Intensité de la mousson
    - Topographie et drainage
    - Gestion des rivières
    - Déforestation
    - Urbanisation
    - Changement climatique
    - Qualité des barrages
    - Envasement
    - Pratiques agricoles
    - Empiètements
    - Préparation inefficace aux catastrophes
    - Systèmes de drainage
    - Vulnérabilité côtière
    - Glissements de terrain
    - Bassins versants
    - Dégradation des infrastructures
    - Indice de population
    - Perte des zones humides
    - Planification inadéquate
    - Facteurs politiques
    - Probabilité d'inondation

- **Source** : Données extraites de Kaggle - Flood Prediction Dataset