Rapport du Projet

MST : Securité IT & Big Data

Prédiction de température avec algorithmes

de machine learning

**Réalisé par :**

**-** AFROUKH Abdellah

**-** BAY BAY Badr

**Encadré par :** Pr. BOUDHIR Anouar Abdelhakim

Année Universitaire : 2024/2025

Table des matières

Introduction

* 1. Contexte et Problématique

Le projet vise à répondre à une problématique liée au changement climatique : **prédire la température moyenne d’un mois donné** en utilisant des données météorologiques historiques et des algorithmes de machine learning. Ce défi est crucial pour des secteurs tels que l’agriculture, l’énergie et les transports. Avec l’essor de l’IA et du ML, il est possible d’exploiter des données climatiques pour effectuer des prédictions précises, mais il reste essentiel de comparer l’efficacité de différents algorithmes afin d’identifier les plus performants pour cette tâche spécifique.

* 1. Objectifs du Projet

1. Construire un modèle prédictif performant pour estimer les températures futures.
2. Comparer les performances de divers algorithmes de régression, comme la régression linéaire et le random forest.
3. Maîtriser le traitement et l’analyse de données météorologiques pour comprendre les tendances climatiques et identifier les variables clés.
4. Effectuer une analyse exploratoire des données (EDA) pour visualiser les relations entre variables et les tendances saisonnières.
5. Automatiser les étapes de prétraitement, de sélection de caractéristiques et d’entraînement dans un pipeline ML.
6. Évaluer les modèles selon des métriques comme **MSE**, **MAE** et **R²**, et optimiser leurs hyperparamètres.
   1. Importance du Projet

Ce projet répond à une problématique actuelle liée au changement climatique et offre des avantages pratiques :

* Amélioration des capacités de prévision pour divers secteurs économiques.
* Comparaison des algorithmes permet une meilleure prise de décision dans des projets similaires.
* Renforcement des compétences en IA et en ML, notamment en collecte, préparation et modélisation des données.

Souhaitez-vous avancer avec la partie suivante sur **la collecte des données** ou avez-vous des remarques sur cette section ?

Collecte des Données

2.1 Sources des Données

Les données météorologiques ont été collectées à partir du site **Open-Meteo**, une alternative fiable qui propose des données climatiques détaillées. Open-Meteo offre un accès à des informations historiques, telles que :

* Températures journalières moyennes.
* Taux d'humidité.
* Précipitations.

2.2 Description des Données Collectées

Les données obtenues via Open-Meteo sont organisées comme suit :

* **Date** : Format ISO (YYYY-MM-DD).
* **Température moyenne** : En degrés Celsius.
* **Humidité** : En pourcentage.
* **Précipitations** : En millimètres.

Ces paramètres sont cruciaux pour comprendre les tendances climatiques et alimenter les algorithmes de machine learning.

2.3 Formatage et Stockage des Données

 **Structuration des données** :

* Les données sont converties et enregistrées dans un fichier **CSV** avec les colonnes : Date, Température (°C), Humidité (%), Précipitations (mm).

 **Nettoyage initial** :

* Suppression des doublons et correction des anomalies dans les valeurs.
* Standardisation des formats et des unités pour garantir la cohérence.

 **Stockage** :

* Les fichiers sont conservés localement ou sur une plateforme cloud pour faciliter l’accès et la collaboration.

Préparation des Données

3.1 Nettoyage des Données

Pour garantir la qualité des données, plusieurs étapes de nettoyage ont été effectuées :

1. **Suppression des doublons** : Toutes les lignes en double ont été retirées.
2. **Gestion des valeurs aberrantes** :
   * Identification des valeurs anormalement élevées ou faibles de température, humidité ou précipitations.
   * Suppression ou correction de ces valeurs à l'aide de méthodes statistiques (ex : moyenne des valeurs proches).
3. **Formatage des dates** : Conversion des dates en un format uniforme (ISO 8601)

3.2 Gestion des Valeurs Manquantes

Les données manquantes ont été gérées de manière systématique :

1. **Analyse des valeurs manquantes** :
   * Calcul du pourcentage de données manquantes par colonne.
2. **Imputation des valeurs** :
   * Pour les températures : Moyenne glissante des valeurs adjacentes.
   * Pour l'humidité et les précipitations : Moyenne ou médiane, selon la distribution.
3. **Suppression** : Si une colonne ou une ligne contenait plus de 30 % de valeurs manquantes, elle a été supprimée.

3.3 Normalisation et Transformation des Données

Les données ont été normalisées pour améliorer les performances des algorithmes :

* **Normalisation Min-Max** : Conversion des valeurs de température, humidité, et précipitations dans une plage de [0, 1].
* **Transformation des dates** : Extraction des caractéristiques comme :
  + Mois (1 pour janvier, 12 pour décembre).
  + Jour de la semaine (1 pour lundi, 7 pour dimanche).

3.4 Extraction des Caractéristiques

De nouvelles caractéristiques ont été ajoutées pour enrichir le dataset :

1. **Moyennes mobiles** : Calcul des moyennes sur des périodes de 7 et 30 jours.
2. **Tendances saisonnières** :
   * Identification des mois les plus froids et les plus chauds.
   * Classification des saisons (hiver, printemps, été, automne).
3. **Interaction des variables** : Création de variables combinées comme Humidité x Précipitations.

Analyse Exploratoire des Données (EDA)

4.1 Analyse des Variables Clés

L’objectif de cette étape est de comprendre les distributions et les tendances des variables principales :

1. **Température moyenne** :
   * Histogramme pour visualiser la répartition des températures.
   * Moyenne et écart-type calculés pour analyser la variabilité.
2. **Humidité** :
   * Boxplot pour détecter les valeurs aberrantes.
   * Analyse des niveaux moyens selon les saisons.
3. **Précipitations** :
   * Histogramme pour examiner la fréquence des jours pluvieux.
   * Analyse des précipitations totales par mois.

4.2 Corrélations entre les Variables

 **Matrice de corrélation** :

* Créée pour visualiser les relations entre les variables clés (Température, Humidité, Précipitations).
* Utilisation de la méthode de corrélation de Pearson.

 **Observations principales** :

* Relation inverse entre Température et Humidité.
* Corrélation modérée entre Précipitations et Humidité.

4.3 Visualisation des Tendances et Saisonnalité

1. **Tendances temporelles** :
   * Ligne temporelle (time series) pour examiner les variations de température au fil des années.
   * Identification des cycles saisonniers (été/hiver).
2. **Tendances saisonnières** :
   * Diagrammes en barres pour les moyennes mensuelles des températures, précipitations et humidité.
3. **Données extrêmes** :
   * Visualisation des jours les plus chauds, les plus humides et les plus pluvieux.

Les graphiques produits dans cette étape fournissent des insights essentiels pour guider la modélisation. Si cela vous convient, je peux continuer avec la section sur **la modélisation**.

Modélisation

5.1 Algorithmes de Machine Learning Utilisés

5.1.1 Régression Linéaire

 **Description** : Un algorithme simple qui modélise la relation entre une variable dépendante (température) et plusieurs variables indépendantes (humidité, précipitations, etc.) en utilisant une ligne droite.

 **Avantages** :

* Rapide à entraîner.
* Facile à interpréter.

 **Limites** :

* Incapable de capturer des relations non linéaires complexes.

5.1.2 Random Forest

 **Description** : Un modèle non linéaire basé sur un ensemble d’arbres de décision, combinant leurs prédictions pour améliorer la précision.

 **Avantages** :

* Capture des relations complexes entre les variables.
* Moins sensible aux valeurs aberrantes.

 **Limites** :

* Plus lent à entraîner.
* Plus difficile à interpréter.

5.2 Division des Données : Entraînement et Test

 Les données ont été divisées en deux ensembles :

* **Ensemble d’entraînement** : 80 % des données, utilisé pour entraîner les modèles.
* **Ensemble de test** : 20 % des données, utilisé pour évaluer les performances des modèles sur des données inconnues.

 **Normalisation appliquée** : Les données d’entrée ont été normalisées pour assurer une convergence optimale des algorithmes.

5.3 Validation Croisée

 **Technique utilisée** : Validation croisée à 5 plis.

 **Objectif** :

* Réduire le surapprentissage (overfitting).
* Obtenir une évaluation robuste de la performance des modèles.

 Les scores obtenus pour chaque pli ont été moyennés pour une meilleure estimation de la précision des algorithmes.

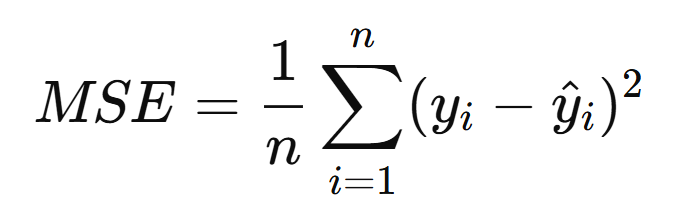
Évaluation des Modèles

6.1 Métriques de Performance

6.1.1 Erreur Quadratique Moyenne (MSE)

 **Description** : Mesure la moyenne des carrés des erreurs entre les valeurs prédites et les valeurs réelles.

 **Formule** :

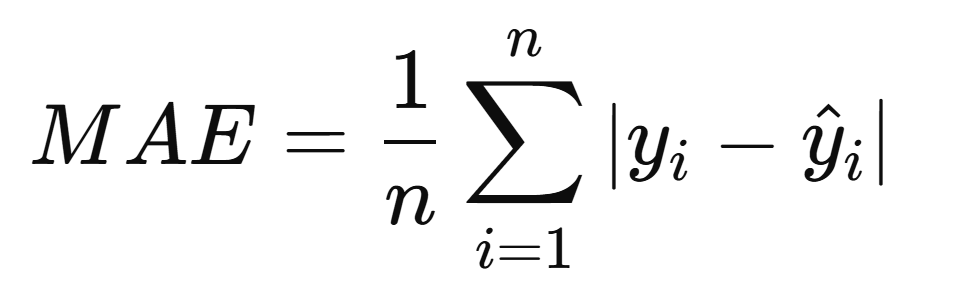


**Interprétation** : Plus la valeur est faible, plus le modèle est précis.

6.1.2 Erreur Absolue Moyenne (MAE)

 **Description** : Calcule la moyenne des écarts absolus entre les valeurs prédites et les valeurs réelles.

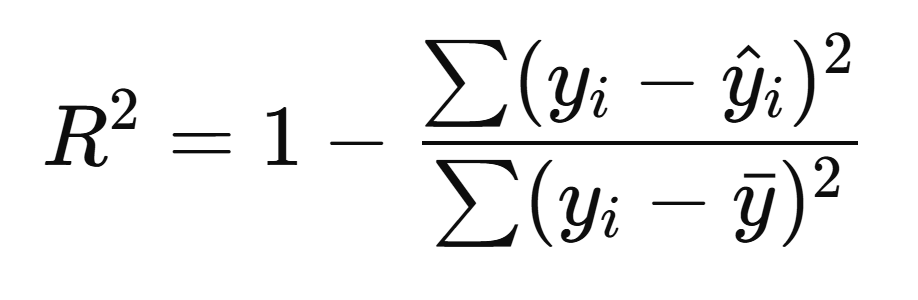
 **Formule** :



* **Interprétation** : Fournit une mesure intuitive des erreurs moyennes en unités des données.

6.1.3 Coefficient de Détermination (R²)

 **Description** : Indique la proportion de la variance des données expliquée par le modèle.

 **Formule** : 

 **Interprétation** :

* R² proche de 1 : Modèle performant.
* R² proche de 0 : Modèle inefficace.

6.2 Comparaison des Modèles

| **Modèle** | **MSE** | **MAE** | **R²** |
| --- | --- | --- | --- |
| Régression Linéaire | Modéré | Faible | Performance moyenne |
| Random Forest | Plus faible que RL | Très faible | Supérieur, capture mieux les relations complexes |

6.3 Forces et Limites des Modèles

 **Régression Linéaire** :

* *Forces* : Rapidité, interprétation facile.
* *Limites* : Faible performance pour des relations non linéaires.

 **Random Forest** :

* *Forces* : Robustesse, capture de la non-linéarité.
* *Limites* : Temps de calcul plus élevé.

Visualisation des Résultats

7.1 Courbes Température Réelle vs Prédite

7.2 Histogrammes des Erreurs

7.3 Comparaison Graphique des Modèles

Technologies et Outils Utilisés

9.1 Bibliothèques Python

Le projet a utilisé plusieurs bibliothèques Python pour la collecte, la préparation, la modélisation et la visualisation des données :

| **Étape** | **Bibliothèque(s)** | **Utilisation** |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |
| **Collecte des données** | requests, pandas | Accéder et structurer les données depuis Open-Meteo. |

|  |  |  |
| --- | --- | --- |
| **Préparation des données** | pandas, numpy | Nettoyage, gestion des valeurs manquantes, normalisation, extraction de caractéristiques. |

|  |  |  |
| --- | --- | --- |
| **Modélisation** | scikit-learn | Entraînement et évaluation des modèles de régression linéaire et Random Forest. |

|  |  |  |
| --- | --- | --- |
| **Visualisation** | matplotlib, seaborn | Création des courbes, histogrammes et graphiques comparatifs. |

9.2 Sources des Données (API ou Datasets)

Les données météorologiques ont été collectées depuis **Open-Meteo**, qui offre des informations fiables sur les températures, humidité et précipitations historiques.

Conclusion

10.1 Résumé des Résultats

10.2 Recommandations pour le Futur

10.3 Limites et Perspectives

Références

Sources des Données

Documentation des Algorithmes

Articles et Livres Utilisés