Breast cancer classifer

X

Weather predictor

Rapport de projet

Réalisé par : Yosri Baati

Année : 2025–2026

1. Introduction Générale

Ce projet explore deux approches essentielles du Machine Learning supervisé : la classification et la régression. L'objectif est de construire un pipeline complet allant de l'exploration des données au déploiement de fonctions de prédiction. Deux jeux de données ont été analysés :

• Le dataset Breast Cancer Wisconsin Diagnostic (classification binaire).

• Un dataset Weather contenant plus de 119 000 enregistrements (régression).

Les étapes étudiées incluent :

– Exploration et visualisation (EDA)

– Prétraitement et nettoyage

– Normalisation

– Entraînement de différents modèles

– Évaluation des performances

– Mise en place de fonctions de test/prédiction.

2. Partie 1 — Classification : Breast Cancer :

2.1. Analyse Exploratoire des Données :

2.1.1. Vue d'ensemble :

- Le dataset contient 569 échantillons et 31 colonnes (30 caractéristiques + 1 variable cible).
- Variable Cible :
  - 0 : Malignant (Maligne) - Cas positif (à détecter)
  - 1 : Benign (Bénigne)
- Types de données : Toutes les caractéristiques sont numériques (float64).

2.1.2. Qualité des Données :

- Valeurs manquantes : Aucune valeur manquante n'a été détectée dans le jeu de données. La base est propre.
- Statistiques Descriptives : Les caractéristiques (rayon moyen, texture, périmètre, aire, etc.) présentent des échelles très différentes (ex: mean area varie de 143 à 2501, tandis que mean smoothness varie de 0.05 à 0.16), ce qui justifie une étape de normalisation.

2.1.3. Visualisation :

Plusieurs graphiques ont été générés pour comprendre la distribution des données :

- Countplot: Visualisation de l'équilibre des classes (légère prédominance des cas bénins).
- Heatmap (Corrélation) : Identification de fortes corrélations entre certaines variables (ex : rayon, périmètre et aire sont très corrélés), ce qui suggère une possible redondance d'information.
- Pairplot & Histogrammes : Analyse des distributions individuelles et des relations entre paires de variables pour distinguer les classes.

2.2. Prétraitement (Preprocessing)

Avant l'entraînement des modèles, les étapes suivantes ont été appliquées :

2.2.1. Séparation des données (Train/Test Split) :

- Le jeu de données a été divisé en un ensemble d'entraînement et un ensemble de test pour évaluer la capacité de généralisation des modèles.

2.2.2. Normalisation (StandardScaler) :

- Utilisation de StandardScaler pour centrer et réduire les variables
- Cela est crucial pour des algorithmes basés sur la distance comme KNN ou SVM, afin d'éviter que les variables à grandes valeurs ne dominent le modèle.

2.3. Modélisation :

Deux algorithmes de classification ont été entraînés et comparés :

- Forêt Aléatoire (Random Forest) : Ensemble d'arbres de décision (Bagging) pour réduire la variance et améliorer la robustesse.
- K-Nearest Neighbors (KNN) : Classifieur basé sur la proximité des voisins, sensible à la normalisation des données.

2.4. Évaluation et Résultats :

Les modèles ont été évalués principalement sur leur Précision (Accuracy), ainsi que sur la Matrice de Confusion et le Rapport de Classification (Précision, Rappel, F1-Score).

Modèle	Précision (Accuracy)
Random Forest	97.66%
K Neighbors	95.90%

Analyse des Résultats:

- Le modèle **Random Forest** a obtenu les meilleurs résultats avec une précision de **97.66%**.
- Le **KNN** suit avec une performance très respectable de **95.90%**.
- La légère supériorité du Random Forest s'explique par sa capacité à gérer des relations non linéaires complexes et sa robustesse intrinsèque (ensemble learning).

2.5. Utilisation dans un contexte médical réel (pour la fonction de test):

- Une telle fonction pourrait être intégrée dans un logiciel d'aide au diagnostic. Par exemple, elle permettrait à un médecin ou à un centre d'imagerie d'obtenir rapidement une prédiction automatique à partir des mesures d'une tumeur. Elle pourrait aussi être utilisée pour un système de triage, aidant à prioriser les cas suspects et à orienter plus rapidement les patients vers un spécialiste. Bien entendu, la prédiction servirait uniquement de support décisionnel, le diagnostic final restant toujours sous la responsabilité du personnel médical.

2.6. Conclusion Partielle:

- Ce projet a démontré qu'il est possible de prédire avec une très haute fiabilité (près de 98%) la nature d'une tumeur du sein à partir de ses caractéristiques.
- Le modèle Random Forest est recommandé pour ce problème en raison de sa performance supérieure (97.66%). Le KNN reste une alternative viable mais légèrement moins précise.

➢ **Pour une application médicale, une attention particulière devrait être portée au Rappel (Recall) afin de minimiser les faux négatifs.**

3. Partie 2 — Régression : Prédiction de la température (Weather Dataset):

3.1. Analyse Exploratoire des Données :

3.1.1. Vue d'ensemble :

- Dimensions : Le dataset est volumineux, contenant 119,040 enregistrements et 31 colonnes.
- Variable Cible : MeanTemp (Température Moyenne).
- Caractéristiques : Comprend MinTemp, MaxTemp, Precip, WindGustSpd, ainsi que des informations de date (Date, Year, Month, Day).

3.1.2. Qualité des Données et Nettoyage :

L'inspection des données a révélé plusieurs défis :

- Valeurs Manquantes : De nombreuses colonnes contenaient un taux élevé de valeurs nulles (ex : WindGustSpd, PoorWeather, Sunshine), nécessitant une gestion appropriée (suppression ou imputation).
- Types de Données : Certaines colonnes numériques étaient mal formatées ou contenaient des valeurs mixtes, nécessitant un nettoyage.
- Sélection de Caractéristiques : Les colonnes peu informatives ou trop incomplètes ont été supprimées pour ne garder que les variables pertinentes (MinTemp, MaxTemp, MeanTemp, Precip, etc.).

3.1.3. Visualisation :

Plusieurs graphiques ont été générés pour comprendre la distribution des données :

- Matrice de Corrélation (Heatmap) :
  - Utilisée pour visualiser l'intensité des liens entre les variables numériques.
  - A révélé une très forte corrélation positive entre MinTemp, MaxTemp et la cible MeanTemp. Cela confirme l'intuition physique : plus les températures minimale et maximale sont élevées, plus la température moyenne l'est.

- Nuages de Points (Scatter Plots) :
  - Ont permis d'observer la linéarité quasi-parfaite entre ces variables de température, validant le choix potentiel d'un modèle de régression linéaire.
- Histogrammes de Distribution :
  - Ont montré la répartition des températures au cours du temps, permettant de vérifier la cohérence des données et l'absence d'aberrations majeures (outliers) après nettoyage.

3.2. Prétraitement (Preprocessing) :

3.2.1. Suppression des colonnes inutiles : Retrait des colonnes totalement vides et des variables non pertinentes pour réduire le bruit et simplifier le modèle.

3.2.2. Nettoyage des données textuelles : Normalisation des colonnes de type string (suppression des espaces, remplacement des virgules, retrait des caractères non numériques) afin de permettre leur conversion en valeurs numériques.

3.2.3. Conversion numérique : Transformation de toutes les colonnes en format numérique, avec conversion automatique des valeurs non valides en NaN.

3.2.4. Séparation Train/Test : Division du jeu de données en un ensemble d'entraînement (80 %) et un ensemble de test (20 %) pour évaluer correctement les modèles.

3.2.5. Imputation : Remplacement des valeurs manquantes par la médiane (calculée uniquement sur l'entraînement) pour conserver un maximum de données.

3.2.6. Normalisation : Application du StandardScaler pour mettre toutes les variables sur la même échelle avant l'entraînement.

3.3. Modélisation :

Deux approches de régression ont été testées :

3.3.1. Régression Linéaire (Linear Regression) : Modèle simple cherchant à établir une relation linéaire entre les variables d'entrée (X) et la cible (y).

3.3.2. Arbre de Décision (Decision Tree Regressor) : Modèle non linéaire capable de capturer des relations plus complexes en divisant l'espace des données en régions.

3.4. Évaluation et Résultats :

Les modèles ont été évalués à l'aide des métriques standards de régression :

- MAE (Mean Absolute Error) : L'erreur moyenne absolue.
- MSE (Mean Squared Error) : La moyenne des erreurs au carré (pénalise les grandes erreurs).
- RMSE (Root Mean Squared Error) : La racine carrée du MSE, dans la même unité que la cible (degrés Celsius).
- R^2 Score : Coefficient de détermination (pourcentage de variance expliquée par le modèle).

Synthèse des Performances :

Modèle	MSE	RMSE	MAE	R^2 Score
Linear Regression	4.31	2.08	1.58	0.94
Decision Tree	3.69	1.92	1.41	0.95

Analyse :

- Les deux modèles ont montré d'excellentes performances pour prédire la température moyenne.
- L'Arbre de Décision obtient des résultats légèrement meilleurs que la Régression Linéaire, avec une erreur plus faible (RMSE de 1.92 contre 2.08) et un coefficient de détermination légèrement supérieur (R^2 de 0.95 contre 0.94).

- Cela indique que, même si la relation entre les variables est largement linéaire, l'Arbre de Décision parvient à capturer certaines non-

linéarités ou variations locales que le modèle linéaire ne peut pas modéliser directement.

Contribution des variables :

Variable	Importance (Arbre de Décision)	Coefficient (Régression Linéaire)
MinTemp	0.995492	8.059178
PRCP	0.004508	-0.316710
PoorWeather	0.000000	0.044510
MO	0.000000	-0.120227

3.5. Utilisation dans un contexte réel (pour la fonction de prédiction météo) :

- Une telle fonction pourrait être intégrée dans une application de prévision météorologique. Par exemple, elle permettrait à une station météo ou à un service en ligne de générer automatiquement la température prévue à partir des mesures enregistrées (température minimale, précipitations, conditions météo, etc.).

  Elle pourrait également être utilisée dans des systèmes automatisés, comme l'agriculture intelligente ou la gestion d'énergie, où des décisions (irrigation, chauffage, ventilation…) reposent sur les conditions climatiques prévues.

- Bien entendu, la prédiction fournie par le modèle servirait uniquement d'outil d'aide, tandis que les décisions finales resteraient entre les mains des opérateurs ou des systèmes métier concernés.

3.6. Conclusion partielle :

- Pour la prédiction de la température moyenne dans ce jeu de données spécifique, l'Arbre de Décision et la Régression Linéaire sont tous deux très performants. L'Arbre de Décision offre une précision légèrement supérieure. La forte corrélation entre les relevés de températures (Min/Max) et la moyenne facilite grandement la prédiction.

4. Conclusion générale :

4.1. Comparison entre classification et regression :

- Dans ce projet, nous avons exploré deux types de problèmes d'apprentissage supervisé :
    - La classification (déterminer si une tumeur est bénigne ou maligne)
    - La régression (prédire une température).

  La classification produit des décisions catégorielles, où l'objectif est d'assigner une observation à une classe. Elle s'évalue avec des métriques comme l'accuracy, la précision ou le rappel.
  La régression, au contraire, prédit une valeur numérique continue, évaluée avec des mesures d'erreur telles que le RMSE, le MAE ou le R^2.
  Bien que les deux approches suivent des étapes similaires (prétraitement, séparation des données, entraînement), leurs objectifs et leurs méthodes d'évaluation diffèrent. La classification cherche à décider « quel type ? », tandis que la régression répond à « quelle valeur ? ».

4.2. Bilan personnel sur les apprentissages :

- Ce projet m'a permis de renforcer ma compréhension des différentes étapes du machine learning : nettoyage et préparation des données, choix des modèles, entraînement, évaluation et interprétation des résultats.
  J'ai également découvert l'importance et **surtout la difficulté** du prétraitement dans la qualité finale des prédictions, ainsi que la différence entre les modèles adaptés aux tâches de classification et ceux spécialisés dans la régression.

- La partie **classification** a été relativement simple : le dataset était de petite taille, propre, sans valeurs aberrantes ni valeurs manquantes. Cela a rendu le prétraitement rapide et l'entraînement du modèle assez direct.

  En revanche, la partie **régression** a été beaucoup plus complexe et a demandé beaucoup plus de temps. Le dataset était volumineux, contenait de nombreuses valeurs manquantes, des colonnes redondantes ainsi que des valeurs incohérentes. Le prétraitement a donc été une étape difficile, marquée par l'incertitude quant aux choix à faire et la crainte d'effacer des informations potentiellement importantes.

  Malgré cela, cette phase m'a permis d'apprendre énormément : techniques d'imputation, stratégies pour traiter des données non numériques, méthodes de nettoyage avancées, et meilleure compréhension du comportement d'un modèle comme le *Decision Tree Regressor*. Ce travail m'a aidé à développer une approche plus rigoureuse et à mieux maîtriser les défis réels liés au traitement de données complexes.

GITHUB REPO