2024

**Mohamed Falilou Fall**

05/09/2024



RAPPORT - PROJET DE MACHINE LEARNING : PREDICTION DES PRIX DE VENTE DE SMARTPHONES (Mission 292)

Table des matières

[1. Introduction 2](#_Toc176458768)

[2. Description des Données 3](#_Toc176458769)

[2.1 Analyse Exploratoire des Données (EDA) 4](#_Toc176458770)

[2.1.1 Corrélation 4](#_Toc176458771)

[2.1.2 Valeurs manquantes et doublons 5](#_Toc176458772)

[2.1.3 Valeurs aberrantes par la visualisation des box plots 5](#_Toc176458773)

[3. Préparation des Données 6](#_Toc176458774)

[4. Modélisation 7](#_Toc176458775)

[4.1 La Régression linéaire 7](#_Toc176458776)

[4.2 Régression Ridge/Lasso 7](#_Toc176458777)

[4.3 Arbre de décision régressif 7](#_Toc176458778)

[4.4 Random Forest (régressif) 7](#_Toc176458779)

[4.5 Random Forest classifier 7](#_Toc176458780)

[4.6 K-Nearest Neighbors (KNN) pour la régression 8](#_Toc176458781)

[4.7 Modèle Sélectionné : Random Forest Classifier 8](#_Toc176458782)

[5. Évaluation du Modèle 8](#_Toc176458783)

[6. Optimisation 8](#_Toc176458784)

[7. Déploiement avec Streamlit 9](#_Toc176458785)

[7.1 Problèmes Rencontrés 9](#_Toc176458786)

[7.2 Le Script ‘app.py’ 10](#_Toc176458787)

[11](#_Toc176458788)

[8. Conclusion 11](#_Toc176458789)

[9. Améliorations Futures 11](#_Toc176458790)

[10. Lien vers les scriptes et l’application : 11](#_Toc176458791)

# 1. Introduction

Le marché des smartphones est en constante évolution, avec une gamme étendue d'appareils aux caractéristiques variées et des prix qui fluctuent en fonction de la technologie, des fonctionnalités et des tendances du marché. Ce projet vise à développer un modèle de Machine Learning capable de prédire avec précision le prix de vente d'un smartphone en se basant sur des caractéristiques spécifiques (comme la taille de l'écran, la capacité de la batterie, la mémoire RAM, etc.).

Les objectifs principaux sont :

- Construire un modèle robuste de prédiction des prix.

- Évaluer les performances du modèle à l'aide de métriques standards.

- Implémenter une application interactive permettant de tester le modèle en temps réel via Streamlit.

# 2. Description des Données

Les données utilisées pour ce projet contiennent des informations sur **292** smartphones, avec les caractéristiques suivantes :

- Taille de l'écran (pouces)

- Capacité de la RAM (Go)

- Capacité de stockage (Go)

- Capacité de la batterie (mAh)

- Résolution de la caméra (MP)

- Prix en $ (variable cible)

2.1 Analyse Exploratoire des Données (EDA)

### 2.1.1 Corrélation

L'analyse exploratoire des données a permis de visualiser les distributions des variables, d'identifier les corrélations, et de détecter les éventuelles valeurs aberrantes. Voici quelques points saillants de l'EDA :

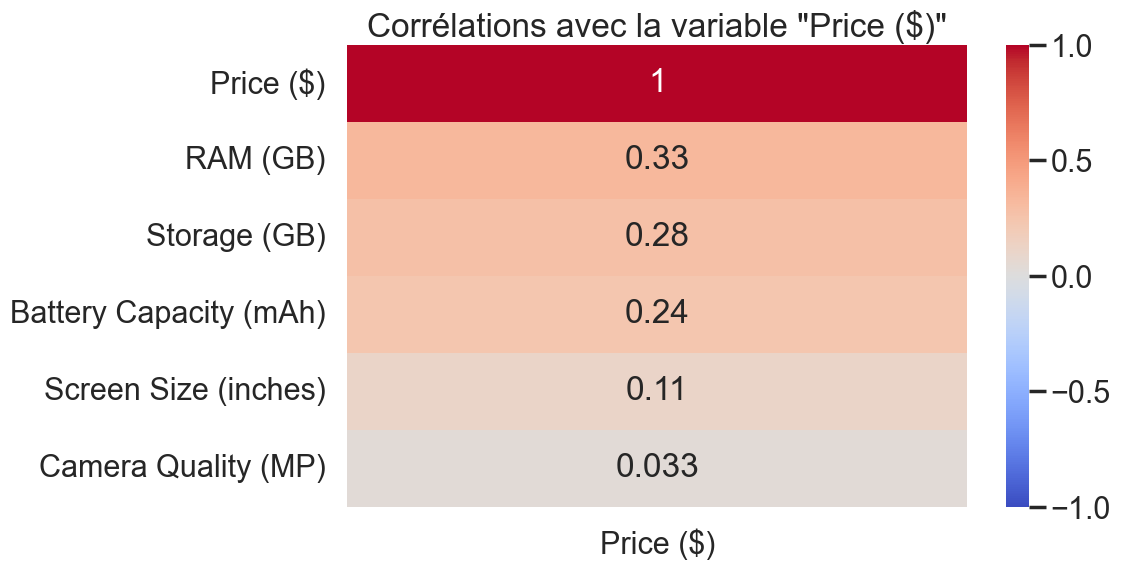
- Corrélation faible entre la capacité de la « RAM (GB) » et le prix soit un degré de relation égal à 33%.

- Corrélation faible entre la capacité de la mémoire interne « Storage (GB) » et le prix soit un degré de relation égal à 28%.

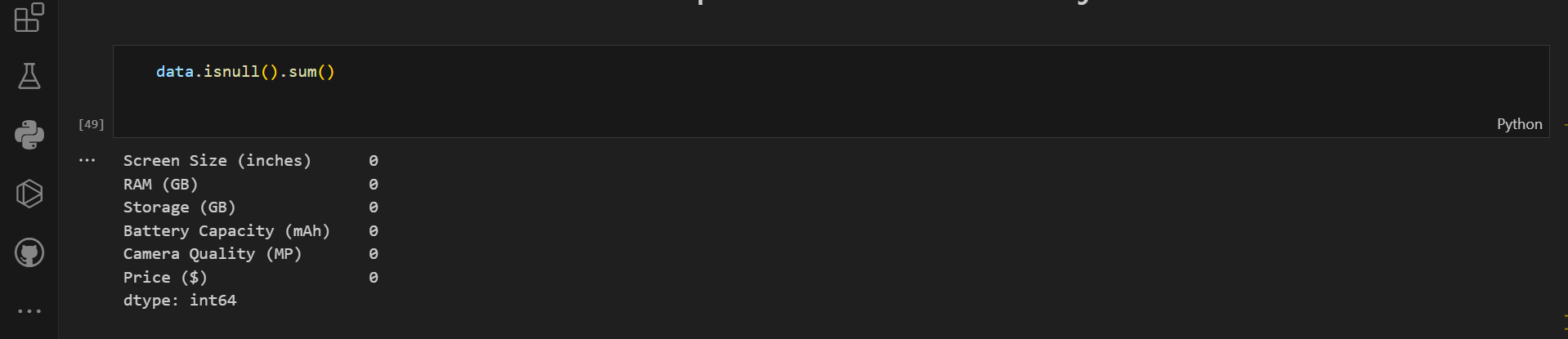
- Corrélation faible entre la puissance de la batterie «Battery Capacity (mAh) » et le prix soit un degré de relation égal à 24%.

- Corrélation faible entre la taille de l’écran en pouces « Screen Size (Inches) » et le prix soit un degré de relation égal à 11%.

- Corrélation faible entre la qualité de la camera en MegaPixel « Camera Quality (MP) » et le prix soit un degré de relation égal à 3,3%.

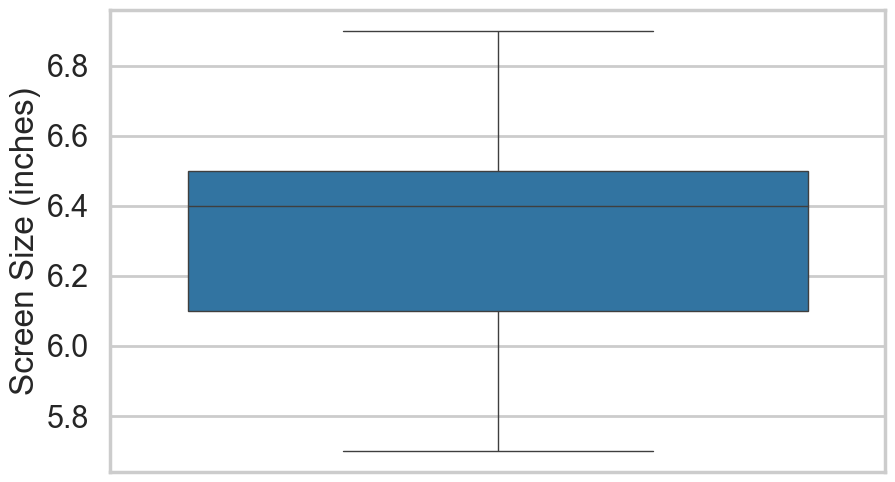
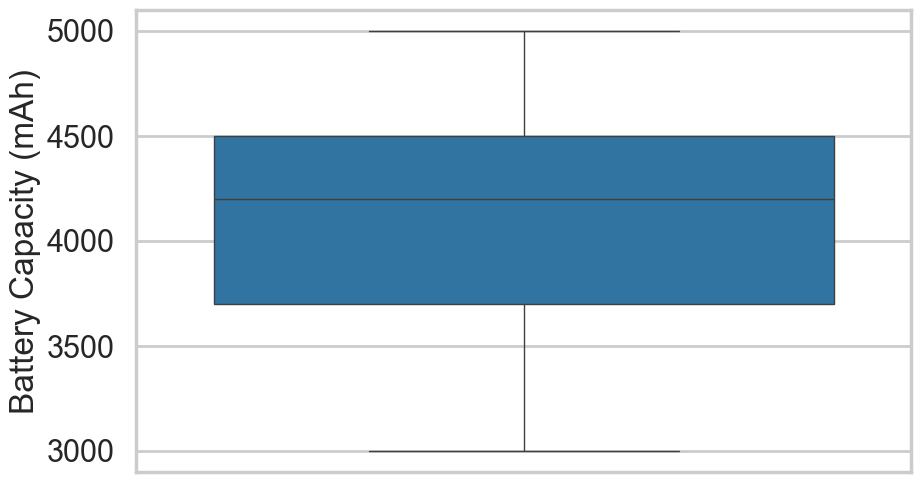


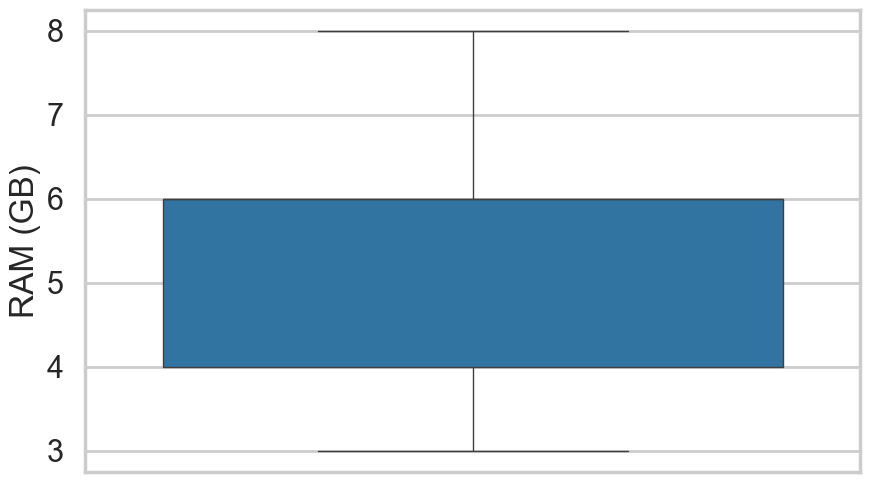
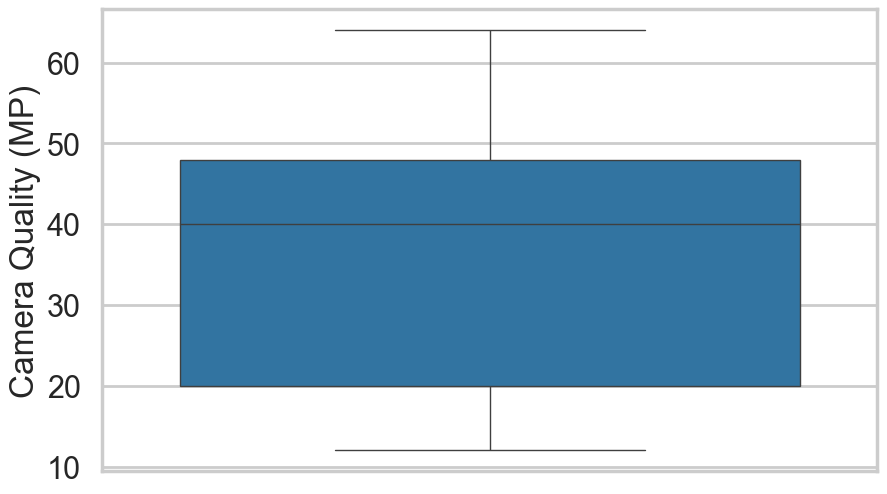
### 2.1.2 Valeurs manquantes et doublons

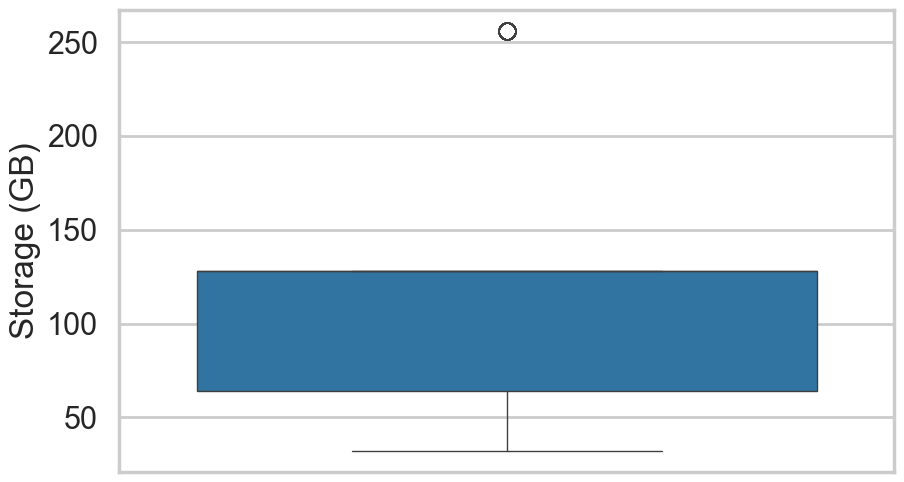
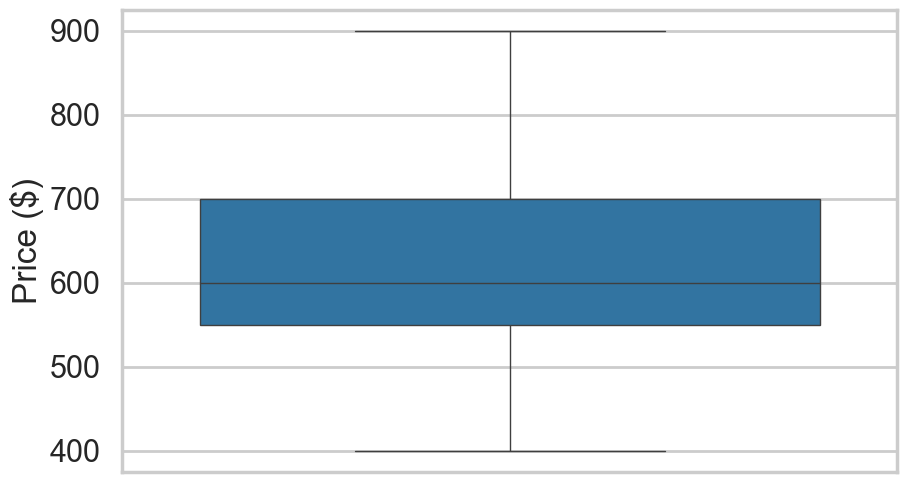


Nous n’avons pas note de valeurs manquantes ni de doublons.

### 2.1.3 Valeurs aberrantes par la visualisation des box plots

# 3. Préparation des Données

Avant d'entraîner le modèle, plusieurs étapes de préparation des données ont été effectuées :

- Nettoyage des données : Les valeurs manquantes, doublons et valeurs aberrantes ont été traités.

- Normalisation : Les caractéristiques numériques ont été normalisées pour s'assurer qu'elles sont sur des échelles comparables

- Séparation des données : Les données ont été divisées en ensemble d'entraînement (70%) et ensemble de test (30%).

# 4. Modélisation

Pour prédire le prix des smartphones, plusieurs modèles de Machine Learning ont été explorés. Les modèles ont été comparés sur plusieurs itérations à l'aide des métriques suivantes : Mean Absolute Error (MAE), Root Mean Square Error (RMSE), et R² Score.

## 4.1 La Régression linéaire

|  |
| --- |
| R2: -2.37323484139874 |
| MAE: 0.7839100373747908 |
| MSE: 0.9275291060310509 |
| RMSE: 0.9630831251927586 |

## 4.2 Régression Ridge/Lasso

|  |
| --- |
| R2: 0.07212340410212115 |
| MAE: 0.7834051444924207 |
| MSE: 0.9260909899853853 |
| RMSE: 0.962336214628435 |

## 4.3 Arbre de décision régressif

|  |
| --- |
| R2: 0.9096677864874954 |
| MAE: 0.07296240097834747 |
| MSE: 0.08915500992934851 |
| RMSE: 0.2985883620125682 |

## 4.4 Random Forest (régressif)

|  |
| --- |
| R2: 0.9133192374894818 |
| MAE: 0.09600993587562602 |
| MSE: 0.07870166606514137 |
| RMSE: 0.28053817220681637 |

## 4.5 Random Forest classifier

|  |
| --- |
| R2: 0.9211465884133401 |
| MAE: 0.09600993587562602 |
| MSE: 0.07870166606514137 |
| RMSE: 0.28053817220681637 |

## 4.6 K-Nearest Neighbors (KNN) pour la régression

|  |
| --- |
| R2: 0.7362419947215917 |
| MAE: 0.19056062373168378 |
| MSE: 0.26325042931865816 |
| RMSE: 0.5130793596692993 |

4.7 Modèle Sélectionné : Random Forest Classifier

Le modèle Random Forest Classifier a été sélectionné comme le plus performant pour cette tâche, car il capture bien les relations non linéaires entre les caractéristiques et le prix cible. Il a également montré une meilleure robustesse face aux données bruitées.

# 5. Évaluation du Modèle

Le modèle Random Forest a été évalué à l'aide de l'ensemble de test. Voici les résultats obtenus :

|  |
| --- |
| R2: 0.9211465884133401 |
| MAE: 0.09600993587562602 |
| MSE: 0.07870166606514137 |
| RMSE: 0.28053817220681637 |

L'analyse des résidus montre que le modèle est capable de bien généraliser sur les données non vues, bien qu'il présente quelques erreurs pour les smartphones dans les gammes de prix extrêmes.

# 6. Optimisation

Des ajustements de paramètres ont été effectués pour optimiser le modèle, notamment en ajustant le nombre d'arbres et la profondeur maximale dans la forêt aléatoire. Une recherche par grille (Grid Search) a permis d'affiner les hyperparamètres, améliorant légèrement les performances.

# 7. Déploiement avec Streamlit

Une application interactive a été développée en utilisant Streamlit pour permettre aux utilisateurs de tester le modèle. L'utilisateur peut entrer les caractéristiques d'un smartphone et obtenir une prédiction de son prix instantanément.

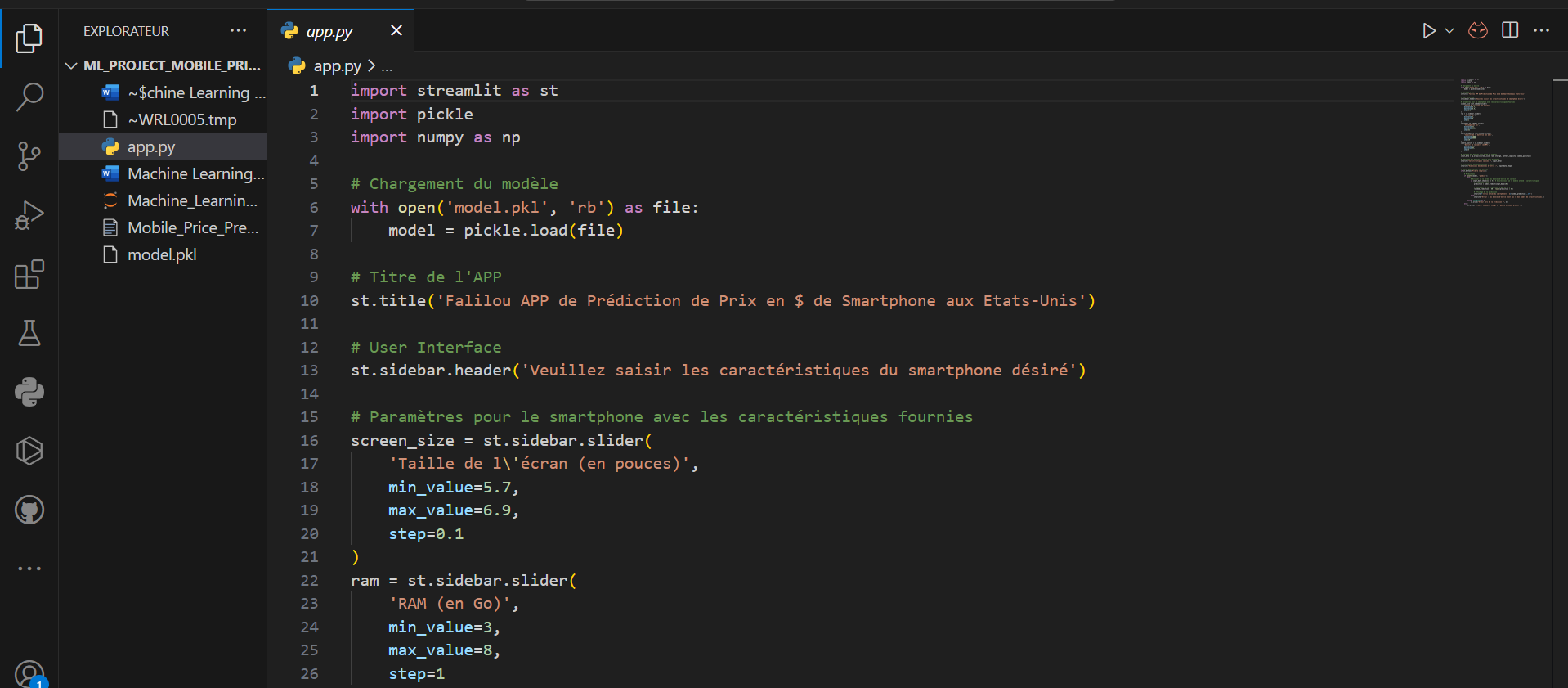


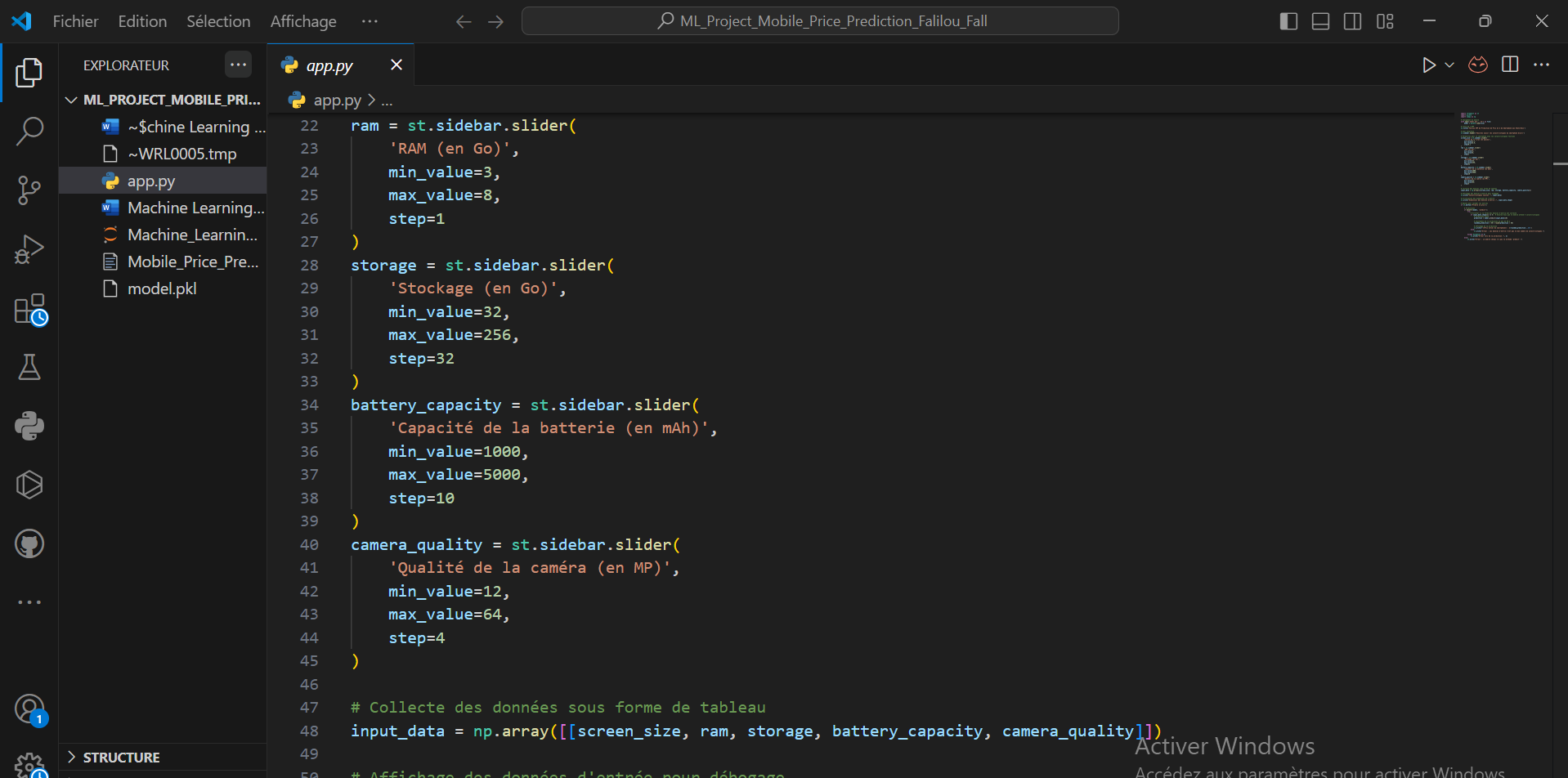
7.1 Problèmes Rencontrés

- Lors de la mise en place initiale de l'application, il a été observé que la prédiction restait fixe, peu importe les modifications des entrées utilisateurs. Après analyse, le problème venait d'une mauvaise gestion des états dans Streamlit. Ce problème a été corrigé en réinitialisant correctement les états à chaque nouvelle entrée utilisateur.

- L'ajustement des plages de valeurs d'entrée a également été optimisé pour s'assurer que le modèle prédit dans des plages de valeurs réalistes.

## 7.2 Le Script ‘app.py’





# C:\Users\hp\OneDrive\Bureau\app3.png

# 8. Conclusion

Le projet a permis de construire un modèle précis et fiable pour la prédiction des prix des smartphones. La combinaison d'une bonne sélection de caractéristiques, d'une modélisation solide (Random Forest Classifier), et d'une application Streamlit fonctionnelle montre un excellent potentiel pour des cas d'utilisation réels dans l'industrie de l'électronique grand public.

# 9. Améliorations Futures

- Incorporation de nouvelles caractéristiques : Ajout d'autres fonctionnalités comme le type de processeur, la qualité des matériaux, la marque et le modèle du Smartphone.

- Une modélisation sur les marques les plus populaires au Senegal.

- Amélioration de l'application Streamlit : Ajout de visualisations plus dynamiques pour montrer l'impact de chaque caractéristique sur le prix prédit.

- Enrichissement des données : Utilisation de données plus vastes et plus variées pour améliorer la généralisation du modèle.

# 10. Lien vers les scriptes et l’application :