simplon
MAGHREB

CAHIER DES CHARGES

FONCTIONNEL ET TECHNIQUE

Sujet de PFE :

Système Intelligent de Détection de Fraude et de Valorisation Douanière (Auto & Tech)

Réalisé par : Mohammed Amine HAMOUTTI

Encadré par : Yassine AMMAMI

Table des Matières

Table des matières

1 Contexte et Problématique

1.1 Contexte : L'Enjeu du Contrôle Douanier

Le contrôle de la valeur transactionnelle des marchandises importées (véhicules d'occasion, matériel informatique) est un défi majeur pour l'administration douanière. La sous-facturation (déclaration d'une valeur inférieure à la réalité) entraîne une perte fiscale significative pour l'État.

1.2 Problématique

Les inspecteurs manquent d'outils automatisés pour vérifier instantanément si la valeur déclarée par un importateur correspond à la réalité du marché marocain. Les bases de données statiques (type Argus) sont souvent obsolètes face à la volatilité des prix (ex : inflation des véhicules d'occasion, décote rapide des smartphones).

1.3 Solution Proposée

Développement d'une plateforme d'**Audit Automatisé** basée sur l'Intelligence Artificielle. Le système compare la *Valeur Déclarée* par l'importateur à une *Valeur de Référence* prédite par un modèle de Machine Learning, afin de lever des alertes de fraude en temps réel.

2 Objectifs du Projet

— **Détection d'Anomalies :** Identifier automatiquement les dossiers suspects présentant un écart injustifié (>20%) entre le prix déclaré et le prix marché.

— **Valorisation Dynamique :** Fournir une estimation "Juste Prix" (Fair Value) basée sur les données réelles du marché marocain.

— **Aide à la Décision :** Visualiser le risque fiscal via un tableau de bord décisionnel (Power BI).

3 Périmètre Fonctionnel (Scope)

Le système traitera trois catégories de produits à fort risque de fraude :

3.1 Module 1 : Véhicules (Cars)

— **Variables d'Entrée :** Marque, Modèle, Année, Kilométrage, Motorisation.

— **Spécificité :** Prise en compte de la dépréciation réelle et de la surcote des modèles populaires (ex : Dacia, Renault).

3.2 Module 2 : Produits High-Tech (Phones & Laptops)

— **Smartphones :** Analyse par Marque, Modèle et Capacité (Stockage).

— **Laptops :** Analyse par Processeur, RAM et Type de Disque (SSD).

4 Spécifications Techniques et Méthodologie

4.1 1. Architecture Data Engineering & Pipeline ETL

La fiabilité du modèle repose sur un pipeline de données robuste et automatisé :

- **Stratégie de Scraping (Extraction) :** Développement de scripts Python (BeautifulSoup/Requests) pour l'extraction massive des données depuis les marketplaces marocaines (ex : Moteur.ma, Avito).
  - Gestion de la pagination et des structures HTML dynamiques.
  - Rotation des User-Agents pour contourner les protections anti-bot.
- **Transformation et Nettoyage (ETL) :** Scripts de pré-traitement avancés pour garantir la qualité des données (Data Quality) :
  - **Nettoyage Regex :** Extraction des valeurs numériques complexes (ex : "120k km" $\rightarrow$ 120000, "1980 ou plus" $\rightarrow$ 1980).
  - **Gestion des Doublons :** Suppression des annonces dupliquées.
  - **Normalisation :** Conversion des devises étrangères en MAD et harmonisation des noms de marques (ex : "VW" $\rightarrow$ "Volkswagen").
- **Simulation de Contrôle (Data Augmentation) :** Génération d'un dataset synthétique de 5000 transactions d'importation via la méthode de **Monte Carlo** (85% conformes / 15% frauduleuses) pour pallier l'absence de données douanières confidentielles.

4.2 2. Modélisation et Évaluation (Machine Learning)

- **Algorithme :** *Random Forest Regressor* (Sélectionné pour sa capacité à capturer les non-linéarités du marché).
- **Feature Engineering :** Vectorisation textuelle (TF-IDF) pour les modèles de véhicules et encodage One-Hot pour les variables catégorielles.
- **Protocole d'Évaluation :** La performance des modèles est mesurée via une validation croisée (Train/Test Split 80/20) sur les métriques suivantes :
  - **MAE (Mean Absolute Error) :** L'erreur moyenne en Dirhams.
  - **R^2 (Coefficient de Détermination) :** Pour évaluer la précision globale de l'ajustement.

4.3 3. Visualisation (Power BI)

Intégration native des scripts Python dans Power BI pour le calcul du "Tax Gap" (Manque à gagner fiscal) et jauge de risque en temps réel.

5 Stack Technologique

- **Langage :** Python 3.12/3.14 (Scripting & ML).

— **Bibliothèques Data :** Pandas (ETL), NumPy, Regex.

— **Machine Learning :** Scikit-learn, Joblib.

— **Visualisation :** Microsoft Power BI.

6 Livrables Techniques

Contrairement à une approche classique, les livrables incluent l'ensemble de la chaîne de valeur de la donnée :

1. **Scripts d'Acquisition (Scraping) :** Code source documenté pour la collecte automatique des données web (`src/scraping/`).
2. **Scripts de Transformation (ETL) :** Pipelines de nettoyage et de standardisation des données brutes (`src/etl/clean_cars.py`).
3. **Modèles Entraînés :** Fichiers binaires sérialisés (`models/car_price_model.pkl`) prêts pour l'inférence.
4. **Datasets Finaux :** Fichiers CSV consolidés et prêts à l'emploi (Données Marché + Simulation Fraude).
5. **Dashboard Power BI :** Rapport interactif (.pbix) intégrant la couche prédictive.