



# **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

<b>1 Contexte et Problématique</b>	<b>2</b>
1.1 Contexte : L'Enjeu du Contrôle Douanier	2
1.2 Problématique	2
1.3 Solution Proposée	2
<b>2 Objectifs du Projet</b>	<b>2</b>
<b>3 Périmètre Fonctionnel (Scope)</b>	<b>2</b>
3.1 Module 1 : Véhicules (Cars)	2
3.2 Module 2 : Produits High-Tech (Phones & Laptops)	2
<b>4 Spécifications Techniques et Méthodologie</b>	<b>3</b>
4.1 1. Architecture Data Engineering & Pipeline ETL	3
4.2 2. Modélisation et Métriques (Machine Learning)	3
4.3 3. Visualisation (Power BI)	3
<b>5 Stack Technologique</b>	<b>3</b>
<b>6 Livrables Techniques et Documentation</b>	<b>4</b>
6.1 1. Code Source (Le Pipeline)	4
6.2 2. Architecture et Documentation	4
6.3 3. Interface et Modèles	4

# 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

Le cœur du projet est un pipeline automatisé garantissant la qualité de la donnée :

- **Module de Scraping (Extraction)** : Scripts Python optimisés pour la collecte massive depuis les marketplaces (Moteur.ma, Avito).
  - Gestion de la pagination et des structures HTML dynamiques.
  - Rotation des User-Agents et gestion des délais pour éviter le blocage.
- **Module ETL (Extract, Transform, Load)** : Scripts de transformation pour assurer l'intégrité des données :
  - **Cleaning** : Nettoyage Regex des champs non structurés (Kilométrage, Année).
  - **Feature Engineering** : Création de variables dérivées (ex : Âge du véhicule).
  - **Loading** : Chargement des données propres dans des fichiers consolidés (CSV/Parquet) prêts pour le ML.
- **Simulation de Contrôle (Data Augmentation)** : Génération d'un dataset synthétique (5000 lignes) via **Monte Carlo** (85% conformes / 15% frauduleuses) pour pallier la confidentialité des données douanières.

### 4.2 2. Modélisation et Métriques (Machine Learning)

- **Algorithme** : *Random Forest Regressor* (Choisi pour sa robustesse aux outliers).
- **Feature Engineering** : Vectorisation TF-IDF pour le texte, StandardScaler pour les numériques.
- **Évaluation de la Performance** : Validation croisée (80/20) mesurée par :
  - **MAE (Mean Absolute Error)** : L'erreur moyenne en Dirhams (Interprétabilité business).
  - **RMSE (Root Mean Squared Error)** : Pour pénaliser fortement les grandes erreurs de prédiction.
  - **R<sup>2</sup> (Score de détermination)** : Pourcentage de variance expliquée par le modèle.

### 4.3 3. Visualisation (Power BI)

Intégration native des scripts Python dans Power BI pour le calcul du "Tax Gap" et le scoring de risque en temps réel.

## 5 Stack Technologique

- **Langage** : Python 3.12 (Scripting & ML).
- **Bibliothèques Data** : Pandas (ETL), NumPy, Regex.

- **Machine Learning** : Scikit-learn, Joblib.
- **Visualisation** : Microsoft Power BI.

## 6 Livrables Techniques et Documentation

Afin de refléter la dimension technique du projet, les livrables sont structurés comme suit :

### 6.1 1. Code Source (Le Pipeline)

- **Scripts de Scraping** : Code documenté pour l'extraction ('src/scraping/').
- **Scripts ETL** : Pipeline de nettoyage et transformation ('src/etl/clean\_data.py').
- **Scripts de Chargement** : Génération des fichiers finaux ('src/etl/load\_data.py').
- **Scripts ML** : Entraînement et évaluation des modèles ('src/models/train.py').

### 6.2 2. Architecture et Documentation

- **Diagramme ETL** : Schéma visuel du flux de données (Source → Transformation → Modèle).
- **Schéma d'Architecture Data** : Vue d'ensemble du système automatisé.
- **Documentation des Données** : Dictionnaire des variables (Data Dictionary) décrivant chaque champ (Source, Type, Transformation appliquée).

### 6.3 3. Interface et Modèles

- **Modèles Sérialisés** : Fichiers '.pkl' prêts à l'emploi.
- **Dashboard Power BI** : Rapport '.pbix' interactif pour l'analyse de fraude.