## 1. ****Contexte****

Pour lutte contre la **contrefaçon de billets, nous avons mis** en place un algorithme capable de **différencier**

**automatiquement les vrais des faux billets** en euros(frnc cfa pour moi), à partir de leurs **caractéristiques géométriques** mesurées par une machine.

.

**Objectif** : Détection de faux billets à partir de mesures physiques (longueur, marges, diagonale, etc.).

1. L’objectif est donc de construire un modèle capable, à partir de ces **mesures**

**géométriques**, de prédire si un billet est **vrai ou faux**.

**Variable cible** : is\_genuine (1 = vrai billet, 0 = faux billet).

## 2. ****Analyse exploratoire & corrélation****

**Répartition** : Countplot montre l’équilibre/déséquilibre entre vrais et faux billets.

**Valeurs manquantes** : Vérifiées avec df.isnull() et heatmap, aucun problème majeur.

**Boxplots** : Pour comparer la distribution des variables selon l’authenticité.

**Matrice de corrélation** :

Visualisée avec sns.heatmap(df.corr(), annot=True, cmap='coolwarm').

Certaines variables (ex. diagonal, height\_right, height\_left) sont fortement corrélées → redondance possible.

Importance : en Machine Learning, forte corrélation peut indiquer multicolinéarité (impact sur certains modèles comme régression logistique, mais moins critique pour Random Forest).

## 3. ****Modèles testés****

**Régression Logistique** (LogisticRegression)

Modèle linéaire, sensible aux variables fortement corrélées et à l’échelle des données.

Utilisation de StandardScaler pour normaliser.

**K-means** (KMeans)

Méthode non supervisée (clustering) → nécessité d’inverser les labels si besoin.

Moins performant car ne tient pas compte de la variable cible.

**KNN** (KNeighborsClassifier)

Algorithme basé sur la proximité entre points.

Dépend fortement de la normalisation et du choix de n\_neighbors.

**Random Forest** (RandomForestClassifier)

Ensemble d’arbres de décision (bagging).

Non sensible à la mise à l’échelle ou à la multicolinéarité.

Capable de gérer variables corrélées et interactions complexes.

Fournit une **importance des variables** pour interprétation.

## ****Pourquoi le choix du Random Forest****

**Robustesse** : Supporte données bruitées et corrélées.

**Précision** : Généralement meilleur score que régression logistique et KNN sur ce type de problème.

**Interprétabilité** : Importance des features pour savoir quelles mesures influencent le plus la décision.

**Flexibilité** : Pas besoin de scaling, fonctionne avec variables de différentes échelles.

**Moins de sur-apprentissage** qu’un seul arbre de décision grâce à l’agrégation de plusieurs arbres

## 5. ****Ce que l’examinateur pourrait te demander****

**Interprétation corrélation** : Pourquoi certaines variables sont corrélées et comment cela influence les modèles.

**Comparaison des modèles** : Forces/faiblesses de chaque méthode testée.

**Pourquoi Random Forest** : Raisons techniques (robustesse, performance, importance des variables).

**Améliorations possibles** : Ajustement hyperparamètres (n\_estimators, max\_depth), validation croisée, feature selection.