PROJET ONCFM : IDENTIFICATION DES CONTREFAÇONS

Analyse et Modélisation

BENMAHAMMED OUSSAMA – DATA ANALYST

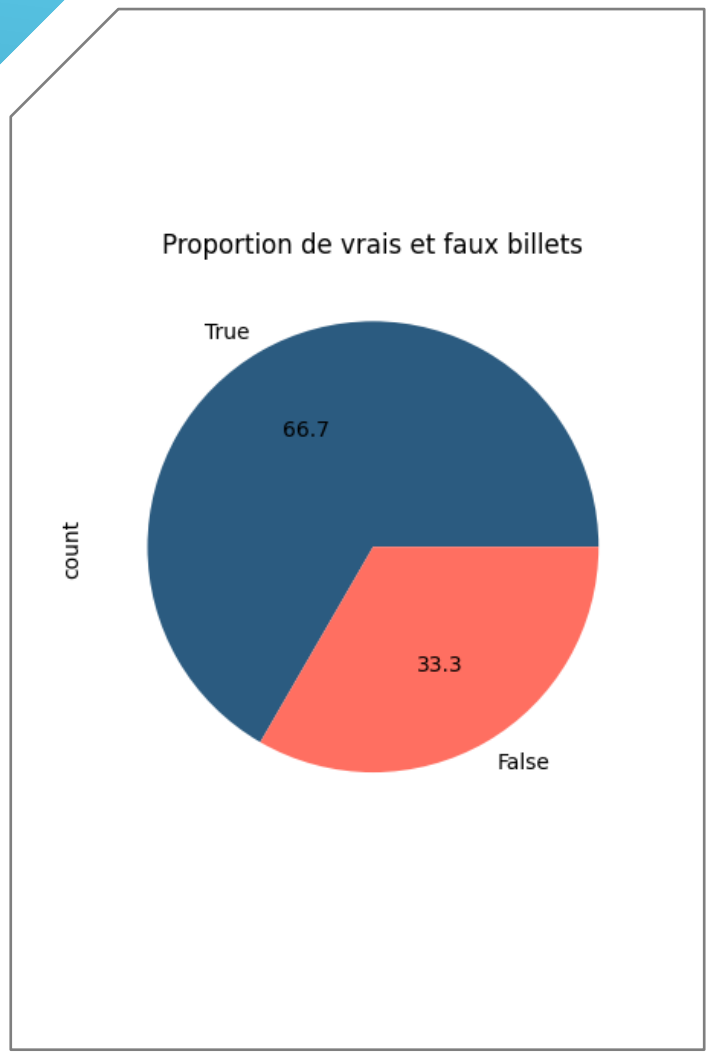
CONTEXTE DU PROJET

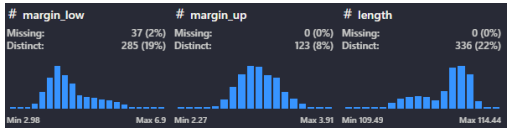
- L'ONCFM souhaite développer un algorithme pour détecter les faux billets. Ce projet vise à explorer différentes méthodes de modélisation.

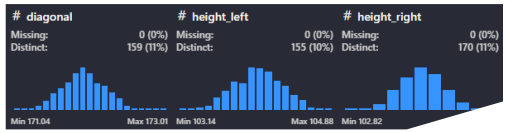
CHEMINEMENT DU PROJET

- 1. Récupération des données
- 3. Prétraitement et analyse exploratoire des données.
- 4. Modélisation avec des techniques supervisées et non supervisées.
- 5. Validation et sélection et optimization du modèle final.

EXPLORATION DES DONNÉES







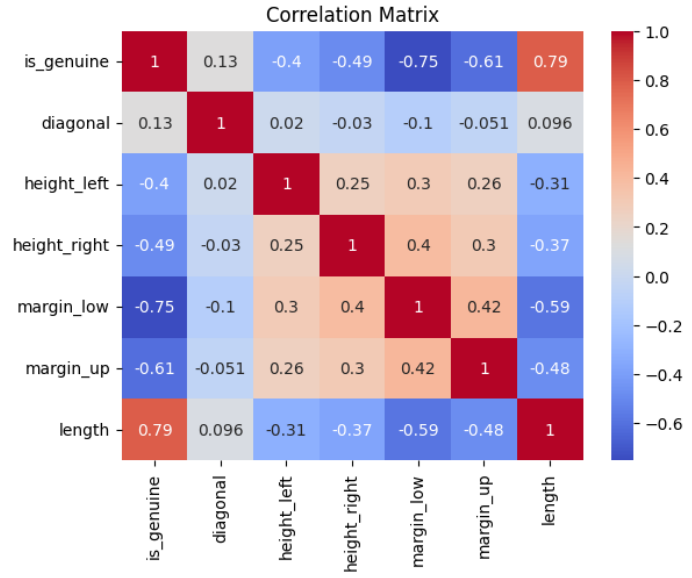
Homogénéité : La majorité des variables, comme diagonal, height_left, height_right, et length, sont bien distribuées, ce qui montre une bonne cohérence dans les mesures des billets.

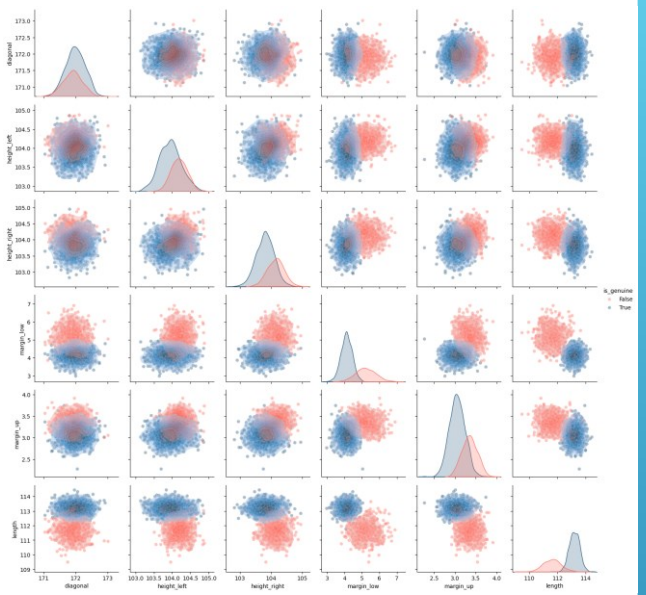
Valeurs Manquantes et Variabilité : margin_low présente des valeurs manquantes et une variabilité plus élevée,

Déséquilibre des Classes : La proportion de vrais et faux billets indique un déséquilibre de classe qu'il faudra gérer lors de l'entraînement des modèles pour éviter des biais.

EXPLORATION DES DONNÉES

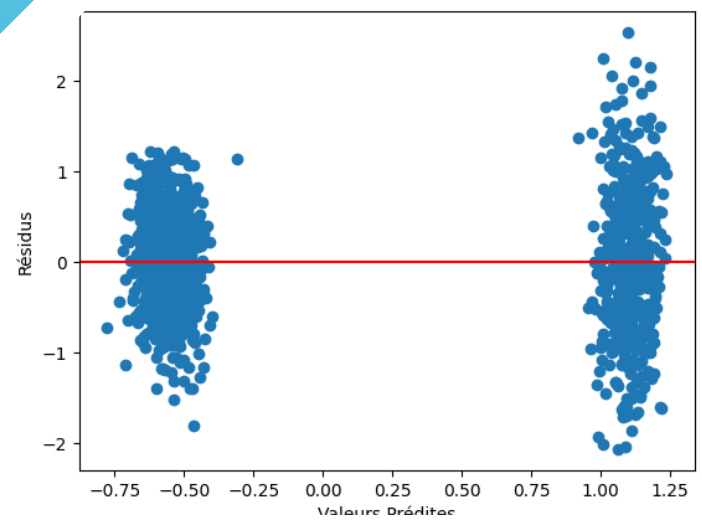
- Relations entre les variables





TRAITEMENTS ET ANALYSES

- Prétraitement des données :
  - Normalisation des données
  - Séparations des données :



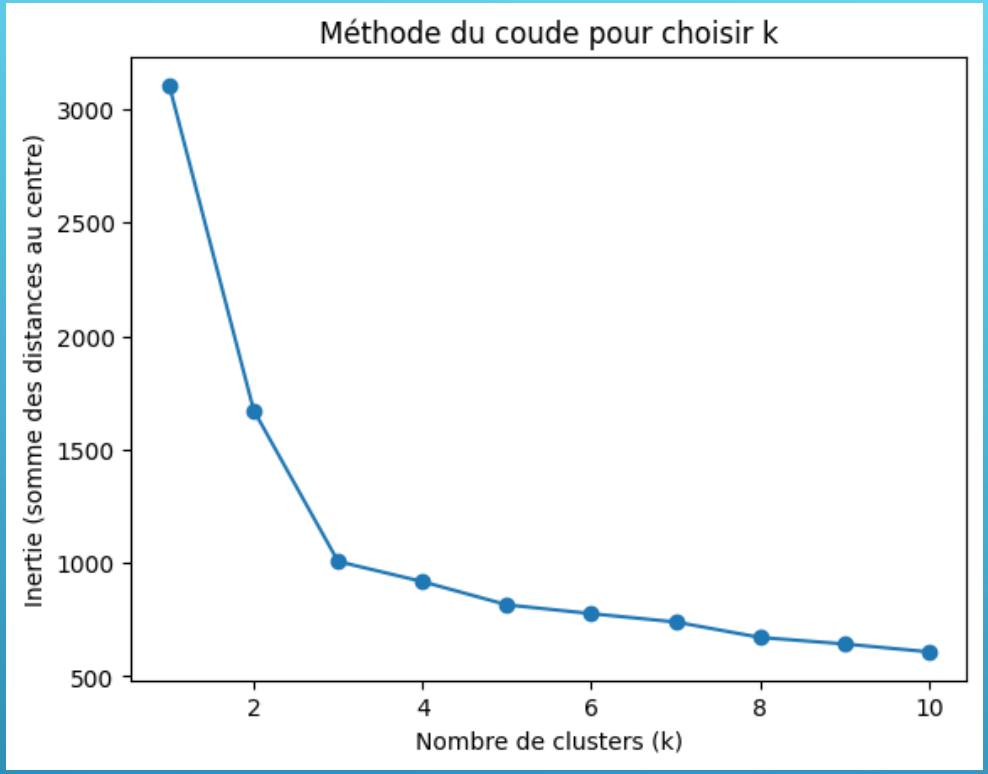
Remplissage des données vides à l'aide de la régression linéaire :

- **R^2** (Coefficient de Détermination) :Le R^2 du modèle est de 0.617, indiquant que 61.7% de la variance de margin_low est expliquée par les variables indépendantes.
- Vérification des Hypothèses de la Régression Linéaire :
  - **a) Linéarité :** Les résidus sont répartis autour de la ligne zéro, indiquant une relation linéaire entre les variables indépendantes et la variable dépendante.
  - **b) Indépendance des Erreurs (Durbin-Watson) :** Valeur de 2.041, confirmant que les résidus ne sont pas autocorrélés.
  - **c) Homoscédasticité des Résidus (Test de Breusch-Pagan) :** La p-valeur est 6.24e-28, indiquant une hétéroscédasticité.
  - **d) Normalité des Résidus (Q-Q plot)**
  - **E) Absence de Multicolinéarité : VIF : 1,46**

TRAITEMENTS ET ANALYSES

MODÉLE NON SUPERVISÉ

- **Modèle K-means :**
  - Avec n_clusters = 2
  - Resultat :
    - Précision : 97%
    - Matrice de confusion

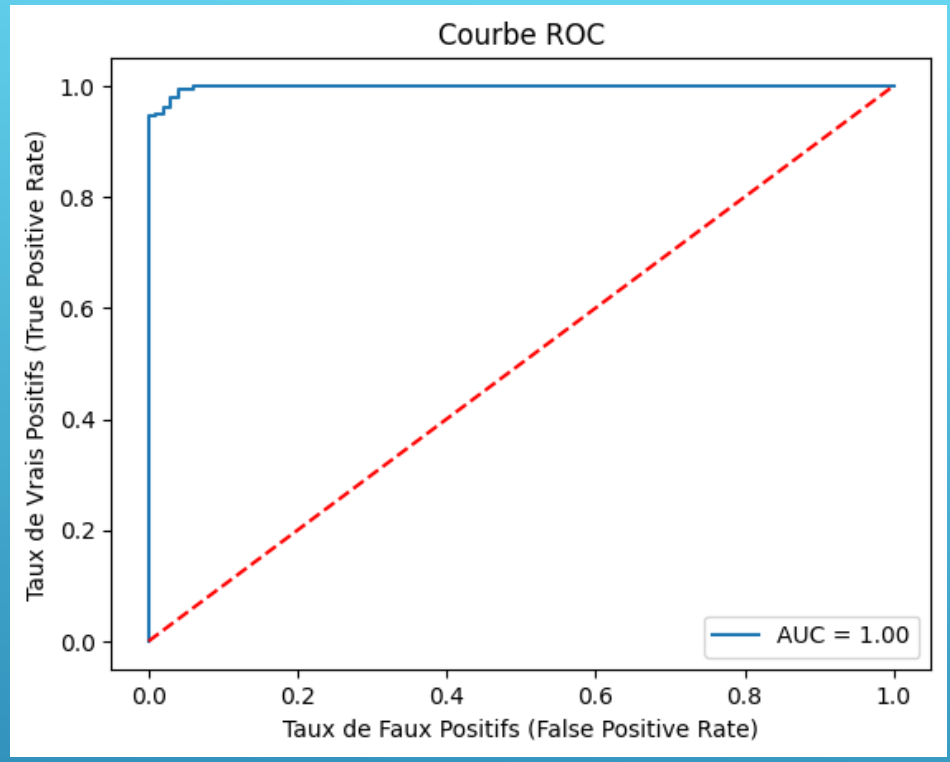


97	3
5	195

MODÉLE SUPERVISÉ

- **Regression logistique :**
  - Resultat :
    - Précision : 98,3%
    - Matrice de confusion

96	**4**
1	**199**

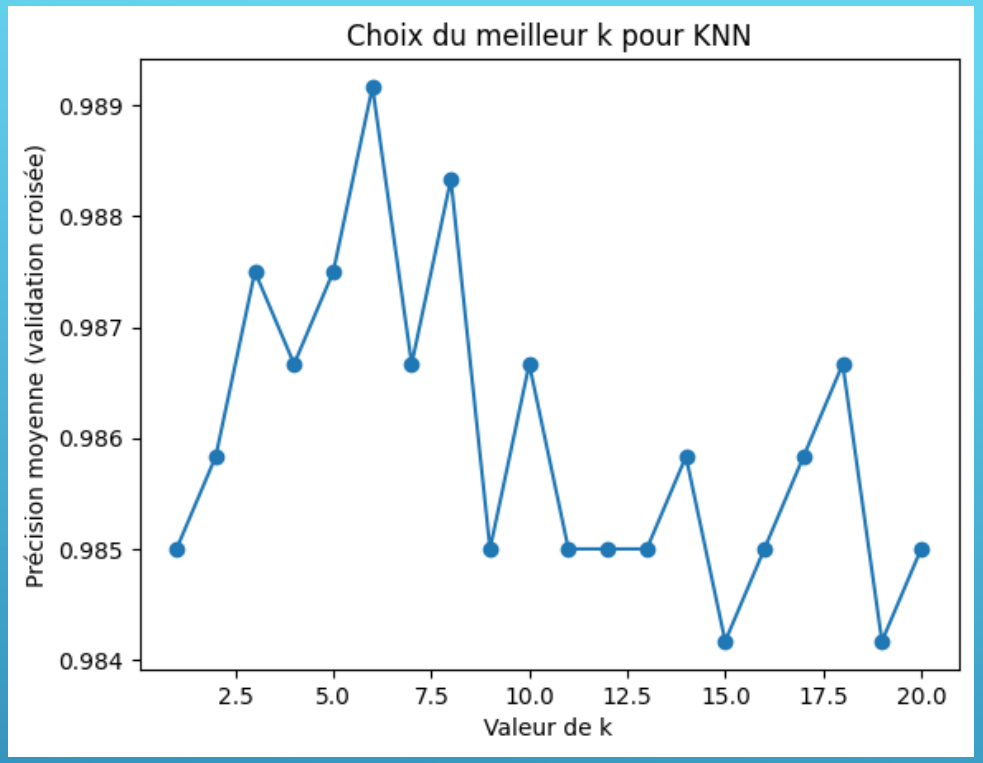


- **Cross-validation**
  - Parametres :
    - n_splits=5
    - shuffle=True
  - **Score moyen : 0.984**

MODÉLISATION

- **KNN**
  - Resultat :
    - Précision : 98,3%
    - Matrice de confusion

96	**4**
1	**199**



- **AvecGridSearch**
  - Parametres :
    - n_splits=5
    - shuffle=True
    - scoring='accuracy'
  - **Score moyen : 0.989**

MODÉLISATION

- **SVM**
  - Resultat :
    - Précision : 99,3%
    - Matrice de confusion

96	4
0	200

Hyperparamètre	Description	Valeurs Possibles	Valeur Optimale
C	Contrôle le trade-off entre une marge large et la classification correcte des points de données	Tout nombre positif (par exemple, 0.01, 1, 10, etc.)	1
gamma	Contrôle l'influence des points d'entraînement individuels pour le kernel non linéaire	'scale', 'auto' ou tout nombre positif (par exemple, 0.1, 1)	'scale'
kernel	Détermine la fonction de transformation des données pour rendre les classes séparables	'linear', 'poly', 'rbf', 'sigmoid'	'rbf'

- **AvecGridSearch**
  - Parametres :
    - n_splits=5
    - shuffle=True
    - scoring='accuracy'
  - **Score moyen : 0.989**

MODÉLISATION

- **RF**
  - Resultat :
    - Précision : 98,6%
    - Matrice de confusion
      - n_estimators=100, max_depth=None

96	4
0	200

Hyperparamètre	Description	Valeurs Possibles	Valeur Optimale
max_depth	Profondeur maximale des arbres, contrôle la profondeur des arbres pour éviter le surapprentissage	Tout nombre positif ou None (illimité)	None
max_features	Nombre de caractéristiques à considérer pour chaque split, contrôle la diversité des arbres	'sqrt', 'log2', ou tout nombre entier positif	'sqrt'
min_samples_leaf	Nombre minimum d'échantillons nécessaires pour être une feuille, évite les feuilles trop petites	1 ou tout nombre entier positif	1
min_samples_split	Nombre minimum d'échantillons nécessaires pour diviser un nœud interne, contrôle la taille des splits	2 ou tout nombre entier positif	2
n_estimators	Nombre d'arbres dans la forêt, détermine la robustesse du modèle	Tout nombre entier positif (par exemple, 50, 100, 200)	50

- **AvecGridSearch**
  - Parametres :
    - n_splits=5
    - shuffle=True
    - scoring='accuracy'
  - **Score moyen : 0.992**

RÉSUMÉ DES RÉSULTATS CLÉS DE L'ANALYSE.

Modèle	Précision (Accuracy)	Rappel (Recall)	F1-Score
Random Forest	0.98	0.96 (False) / 0.99 (True)	0.97 (False) / 0.99 (True)
SVM	0.99	0.96 (False) / 1.00 (True)	0.98 (False) / 0.99 (True)
KNN	0.98	0.96 (False) / 0.99 (True)	0.97 (False) / 0.99 (True)
Régression Logistique	0.98	0.96 (False) / 0.99 (True)	0.97 (False) / 0.99 (True)
K-means	0.97	0.97 (False) / 0.97 (True)	0.96 (False) / 0.98 (True)

RÉSULTATS

- Le modèle final retenu est celui qui a obtenu le meilleur score :

“RANDOM FOREST “