DATAMINING

Mini PRojet

RÉALISÉ PAR:
AYCHA CHOUCHENE

20-10-2025

SOMMAIRE

PRÉSENTATION DU DATASET

INFORMATIONS GÉNÉRALES

NOM : ELLIPTIC BITCOIN DATASET
SOURCE : KAGGLE
NOMBRE DE TRANSACTIONS : 203 769
NOMBRE DE FEATURES : 166
NOMBRE DE CLASSES: 3
TYPES DE DONNÉES : NUMÉRIQUES
CLASSES : 1:LICIT, 2:ILLICIT, 3:UNKNOWN
UTILISATION : DÉTECTION DE FRAUDE ET BLANCHIMENT D'ARGENT, DONNÉES ANONYMISÉES

DATA PREPARATION

fusion.ipynb

- Chargement des fichiers CSV
- Fusion des features et labels sur txId
- renommage de features (feature_1, feature_2, ...)
- Export du dataset fusionné pour l'EDA

```
data = pd.merge(features, labels, on='txId')
```

EXPLORATORY DATA ANALYSIS

Définition

L'EDA CONSISTE À EXAMINER ET COMPRENDRE LES DONNÉES AVANT TOUTE MODÉLISATION

Objectifs principaux

- VÉRIFIER LES VALEURS MANQUANTES ET DOUBLONS
- ÉTUDIER LA RÉPARTITION DE LA VARIABLE CIBLE (TARGET ANALYSIS)
- EXAMINER LES FEATURES (DISTRIBUTION, OUTLIERS, CORRÉLATIONS)

EXPLORATORY DATA ANALYSIS

Valeurs manquantes et doublons

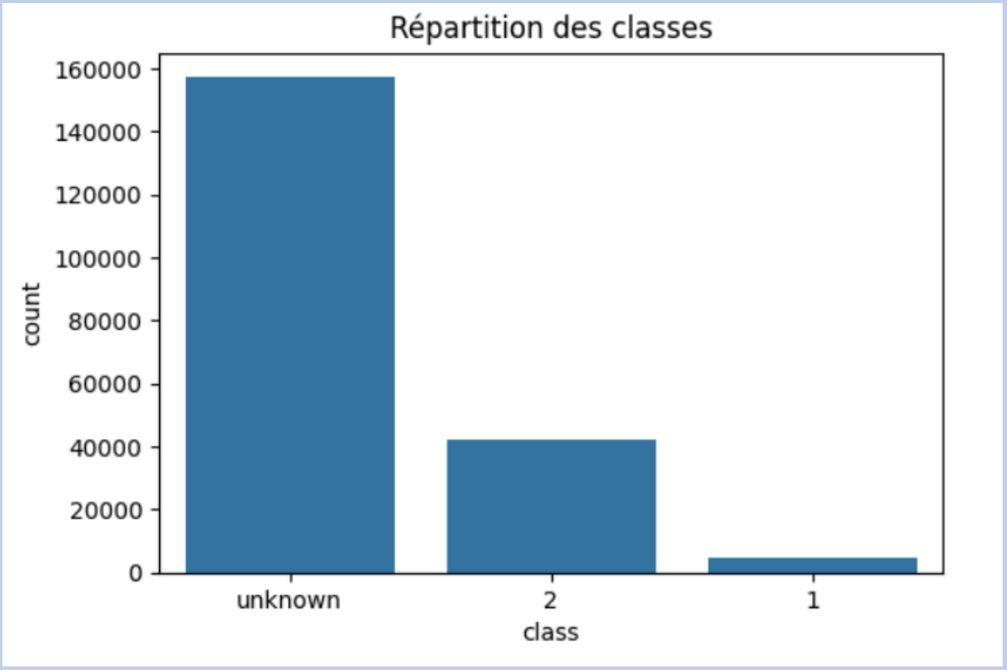
- DÉJA GÉRÉS

```
Nombre de doublons : 0
```

```
Valeurs manquantes par colonne :
txId           0
feature_115    0
feature_107    0
feature_108    0
feature_109    0
```

Target Analysis

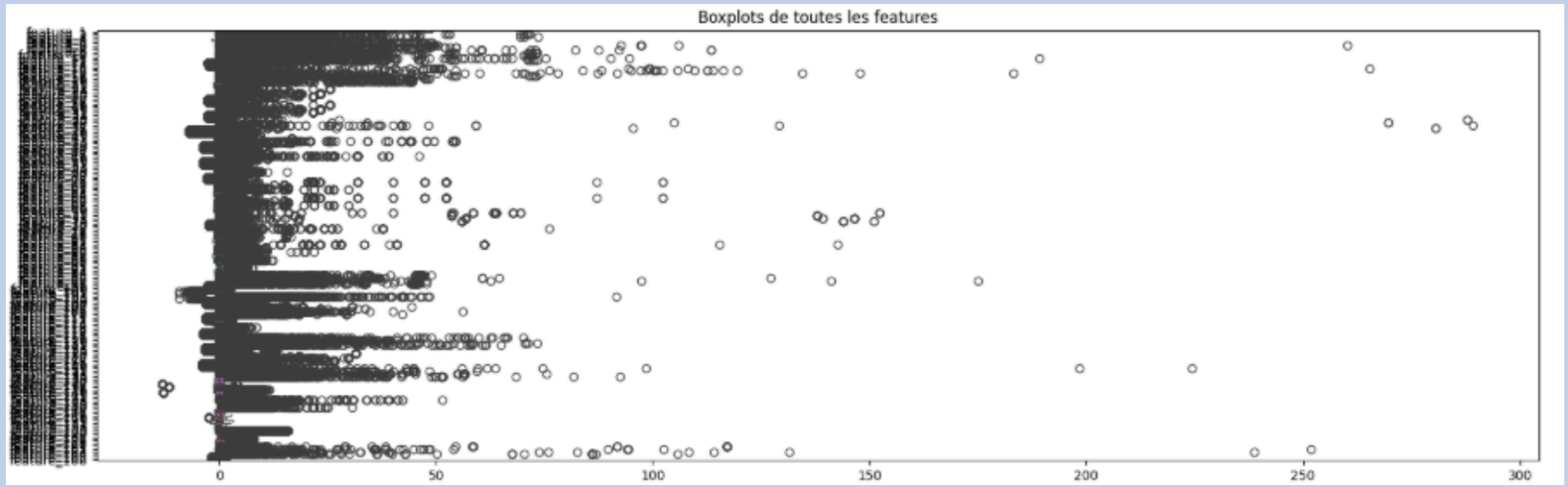
- RÉPARTITION DES CLASSES
- VISUALISATION PAR BARPLOT

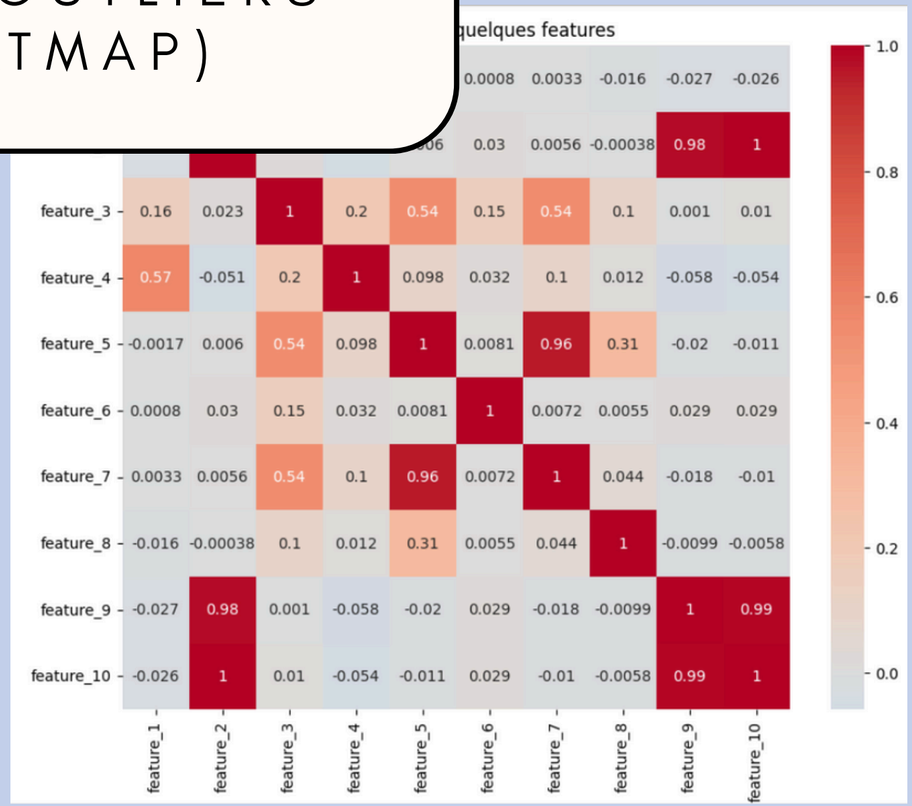


EDA

Feature analysis

- HISTOGRAMMES ET BOXPLOTS POUR VISUALISER LA DISTRIBUTION ET LES OUTLIERS
- CORRÉLATIONS ENTRE FEATURES (HEATMAP)





FEATURE ENGINEERING

Définition

LE FEATURE ENGINEERING CONSISTE À CRÉER, SÉLECTIONNER ET TRANSFORMER LES VARIABLES POUR AMÉLIORER LA QUALITÉ DES DONNÉES ET LA PERFORMANCE DES MODÈLES

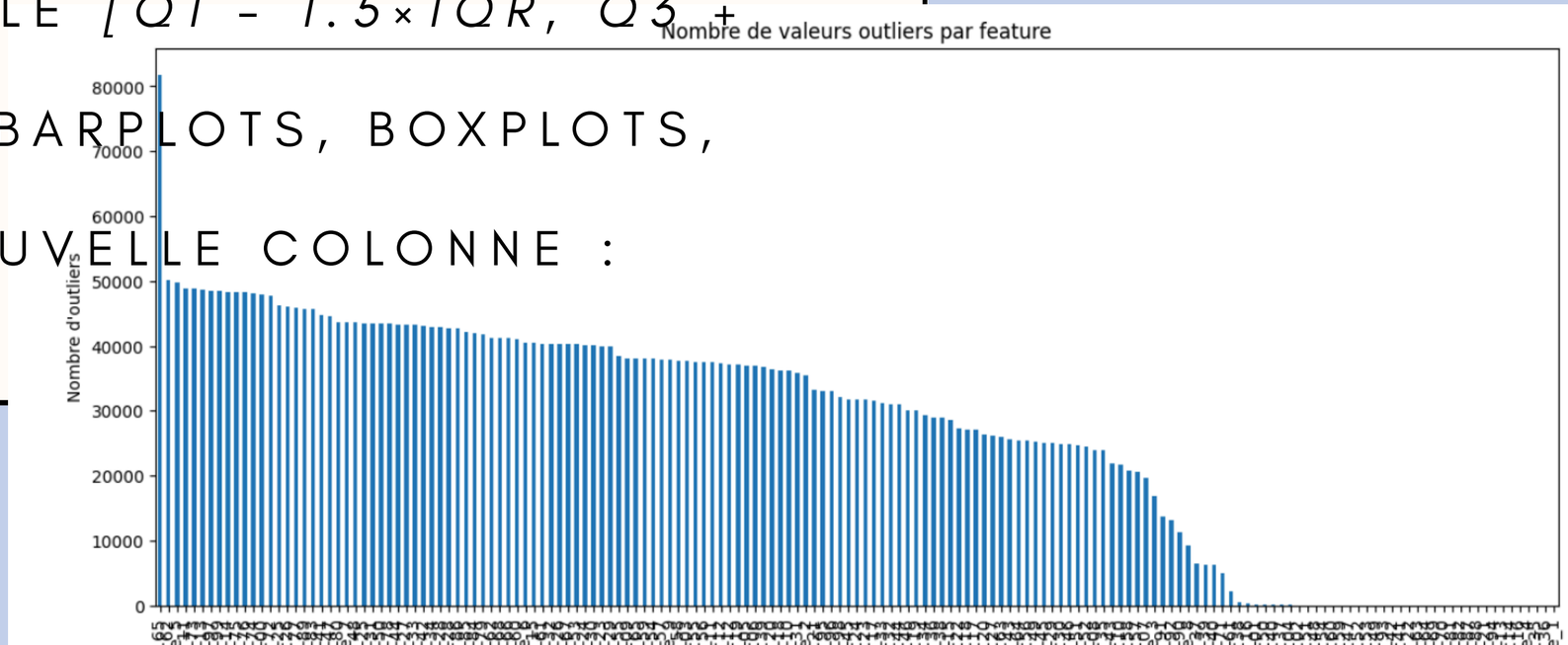
Démarche appliquée

1. DÉTECTION ET TRAITEMENT DES OUTLIERS
2. SÉLECTION INITIALE DE FEATURES
3. TRANSFORMATION DES DONNÉES
4. CONSTRUCTION DU DATASET FINAL

DÉTECTION ET TRAITEMENT DES OUTLIERS

1. Détection univariée : IQR (Interquartile Range)

- ANALYSE VARIABLE PAR VARIABLE.
- PERMET D'IDENTIFIER LES VALEURS SITUÉES HORS DE L'INTERVALLE *[Q1 - 1.5×IQR, Q3 + 1.5×IQR]*.
- VISUALISATION PAR BARPLOTS, BOXPLOTS, HEATMAP
- CRÉATION D'UNE NOUVELLE COLONNE : *IS_OUTLIER_UNIV*

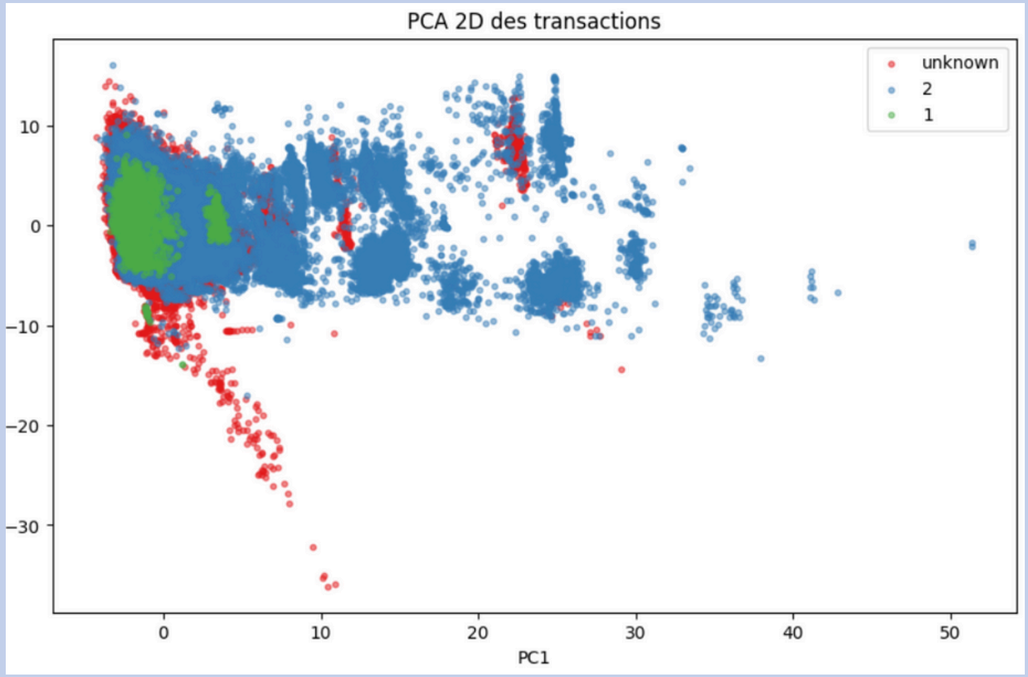


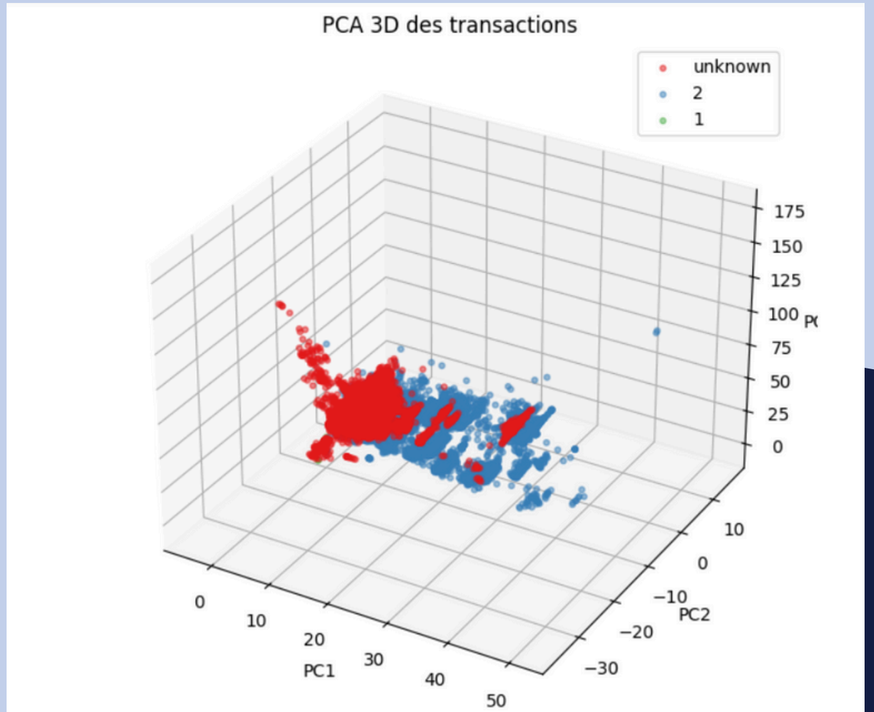
DÉTECTION ET TRAITEMENT DES OUTLIERS

2. Détection multivariée : Isolation Forest

- Méthode basée sur les arbres de décision.
- Isole les observations “anormales” plus rapidement que les normales.
- Nouvelle colonne : is_outlier_multiv.
- Visualisation : PCA 2D/ 3D

DANS LE DATASET ELLIPTIC, LES OUTLIERS PEUVENT REPRÉSENTER DES COMPORTEMENTS INHABITUELS DE TRANSACTIONS, DONC POTENTIELLEMENT FRAUDULEUX.





SÉLECTION INITIALE DE FEATURES

Objectif

Réduire le nombre de variables tout en conservant celles qui apportent le plus d'information utile pour la détection de fraude.

Méthodes utilisées

1. Filtrage par variance
2. Analyse de corrélation
3. Information mutuelle

le nombre de features est passé de 166 → 87.

TRANSFORMATION DES DONNÉES

Objectif

Mettre toutes les variables à la même échelle afin d'éviter qu'une feature à grande amplitude domine les autres lors de l'entraînement du modèle.

Méthodes appliquées

Standardisation (StandardScaler)

Transforme les données pour avoir :

une moyenne = 0

un écart-type = 1

Normalisation (MinMaxScaler)

Ramène les valeurs dans un intervalle [0, 1].

CONSTRUCTION DU DATASET FINAL

Étapes principales

1. Ajouter la colonne class
2. Ajouter les colonnes d’outliers
3. Ajouter txld
4. Exporter le dataset final

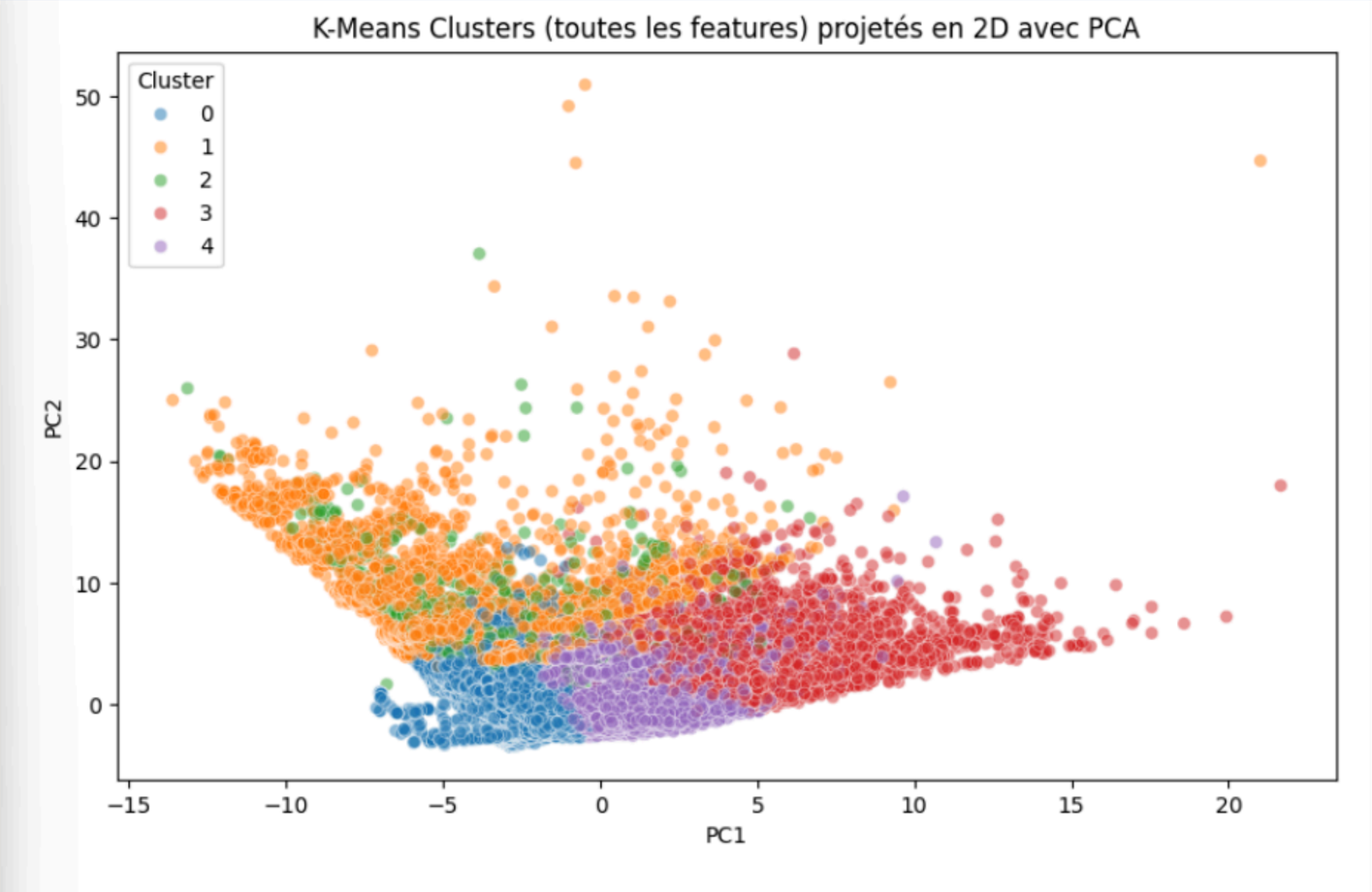
MODÉLISATION NON SUPERVISÉE

Étapes principales :

1. Déterminer le meilleur nombre de clusters (k): Elbow Method, Silhouette Score
2. K-Means sur les features numériques (91 features)
3. Réduction de dimension avec PCA
4. K-Means sur les données réduites
5. Evaluation et comparaison

K-MEANS SUR LES FEATURES NUMÉRIQUES

```
Inertia: 16040299.83, Silhouette Score: 0.068
```

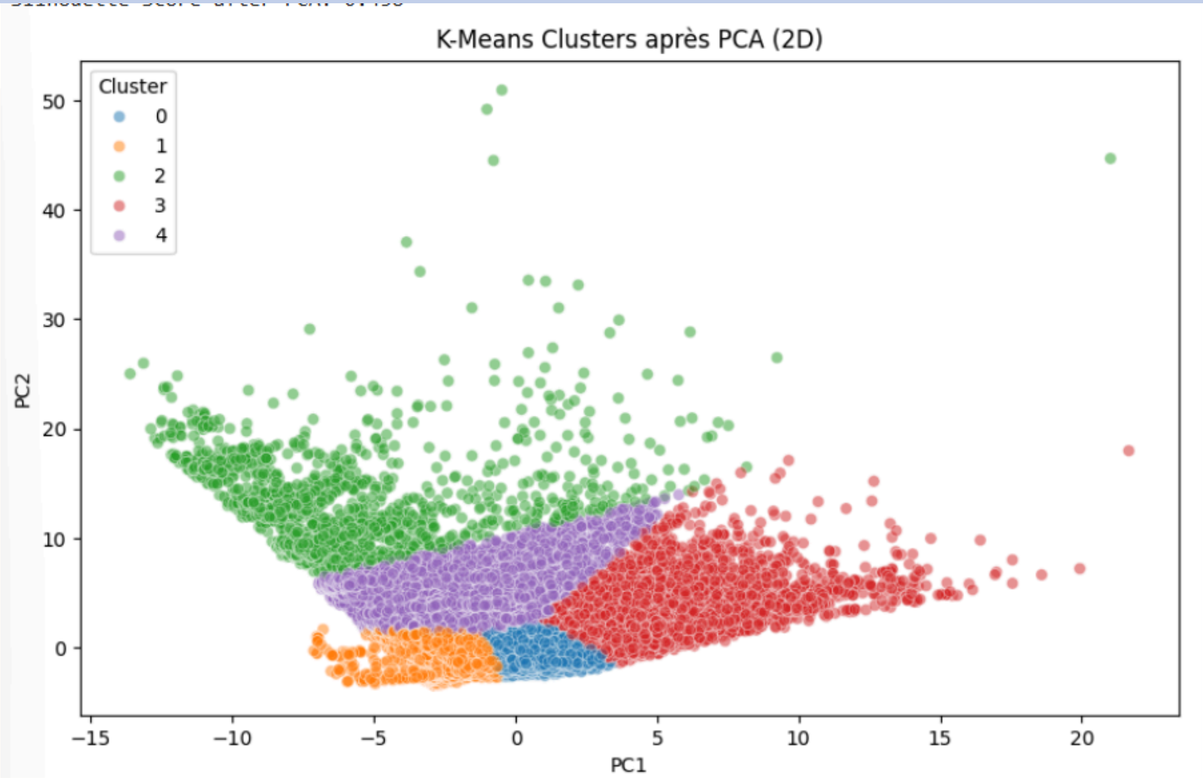


Inertie élevée :
Les points sont éloignés de leurs centres de clusters,
Ce qui traduit une structure floue et continue.
Les clusters se fondent les uns dans les autres au lieu d'être clairement séparés

Silhouette Score faible (0,068) :
La séparation entre les clusters est très faible, confirmant un fort chevauchement et une définition des clusters peu claire.

K-MEANS APRÈS RÉDUCTION DE DIMENSION PCA

```
Silhouette Score after PCA: 0.438
```

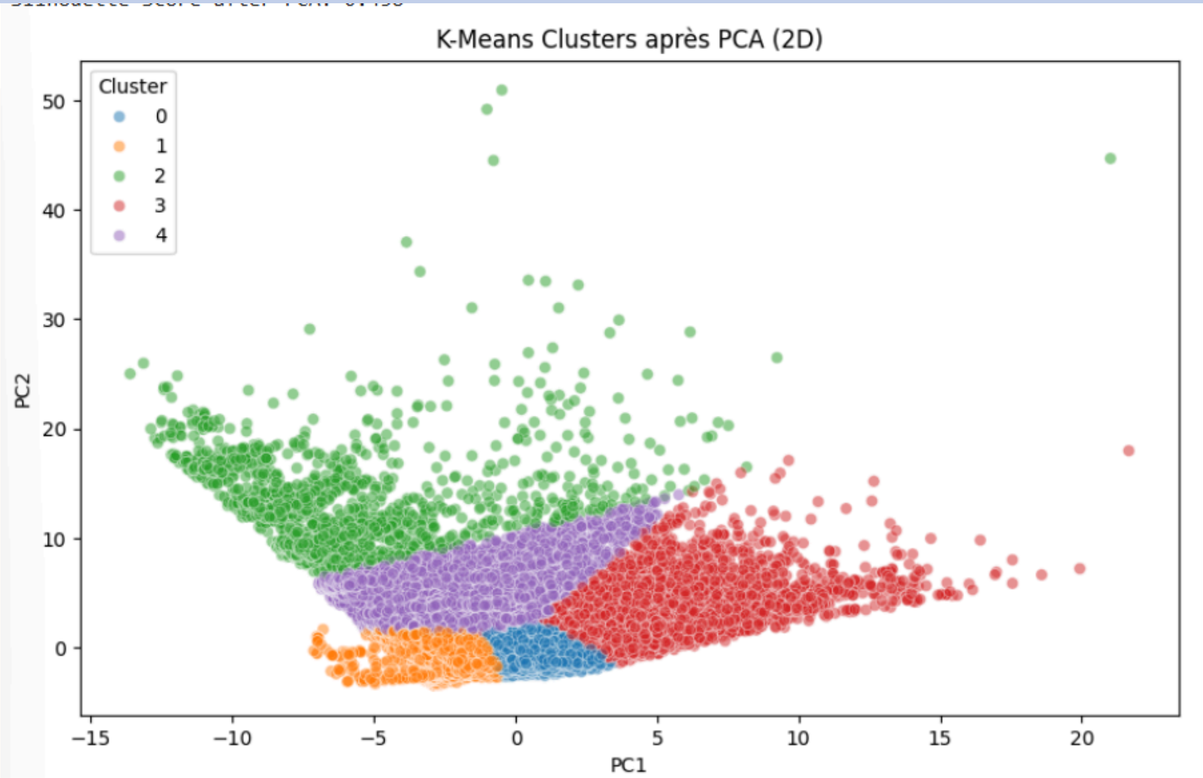


Silhouette Score = 0,438 :
La qualité du clustering est modérée à **bonne**, ce qui montre que la réduction de dimension avec PCA a permis de mettre en évidence des motifs intéressants dans les transactions.

Observation des clusters :
Les groupes sont plus distincts qu'avant PCA

VISUALISATION DES CLUSTERS AVEC OUTLIERS

```
Silhouette Score after PCA: 0.438
```

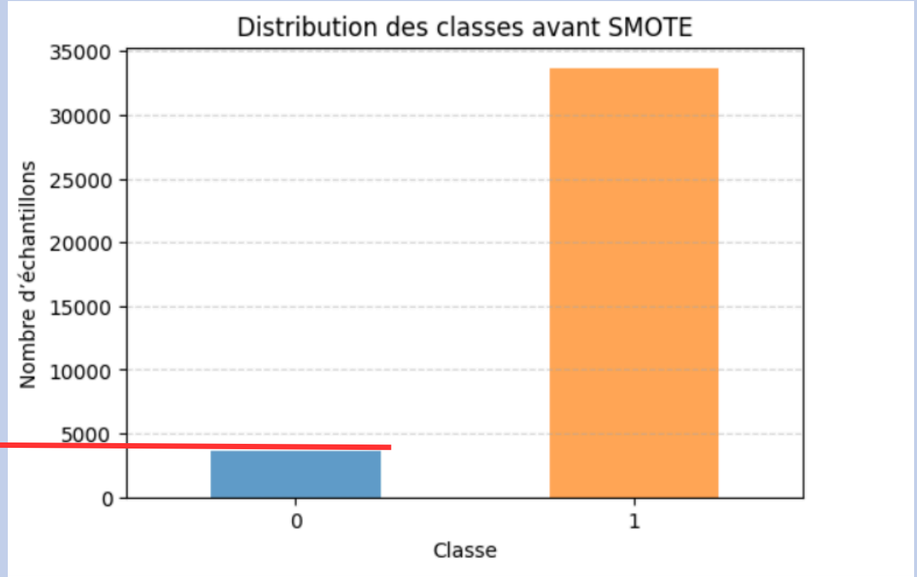


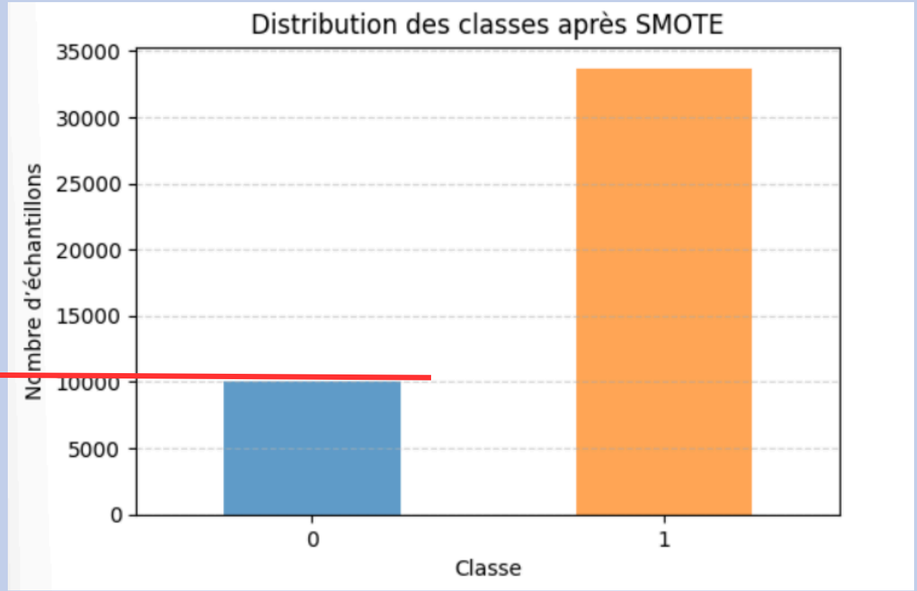
Les points isolés peuvent également indiquer une activité suspecte, car ils se distinguent des transactions denses et typiques.

MODÉLISATION SUPERVISÉE

Préparation des données

1. NETTOYAGE DES LABELS
2. ENCODAGE DE LA TARGET
3. SÉPARATION TRAIN/TEST
4. GESTION DU DÉSÉQUILIBRE: **SMOTE**



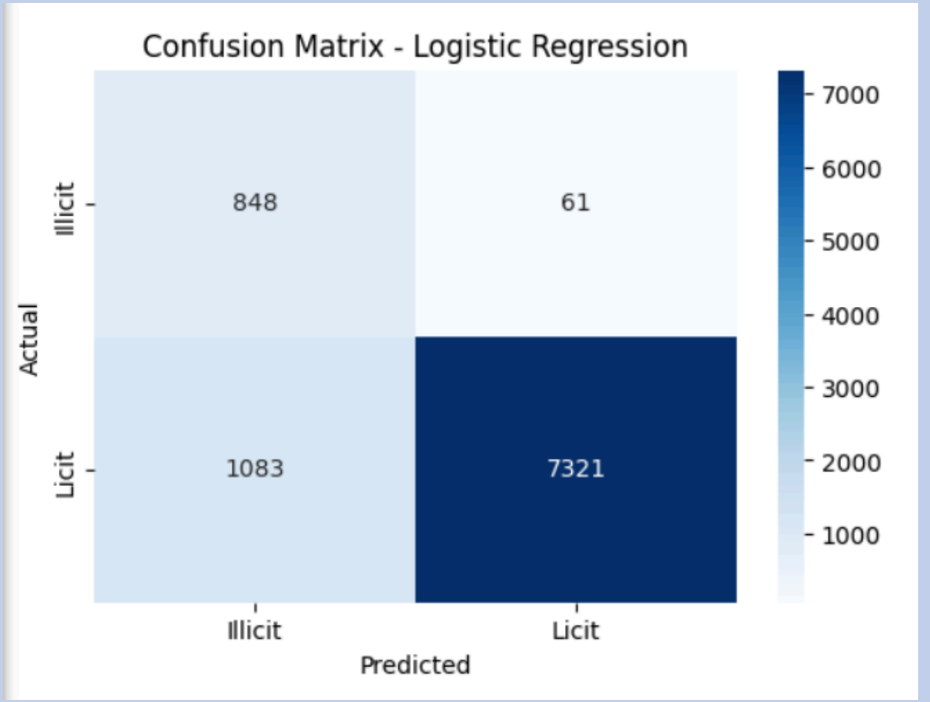


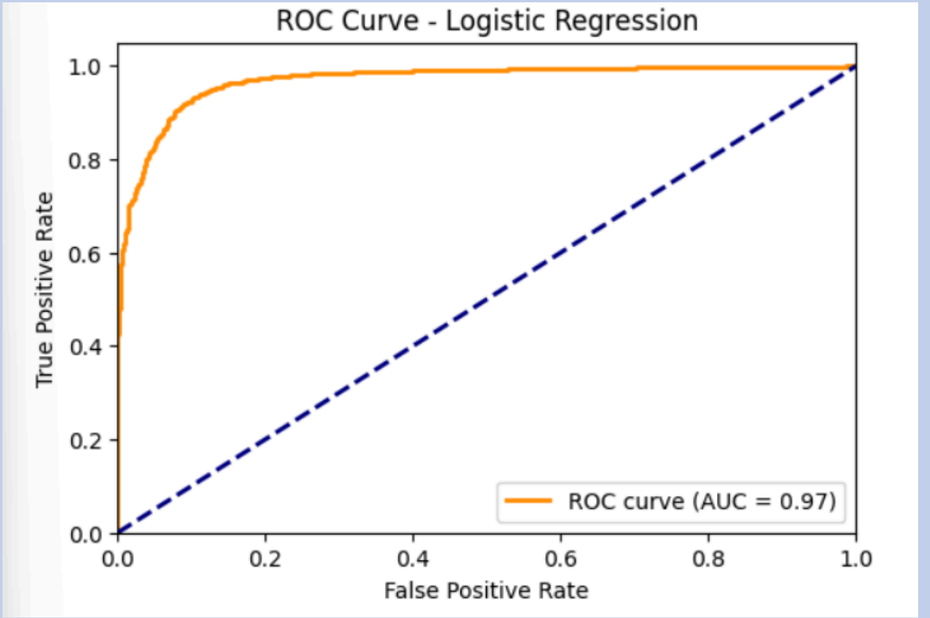
LOGISTIC REGRESSION

MODÈLE LINÉAIRE QUI PRÉDIT LA PROBABILITÉ QU'UNE TRANSACTION SOIT FRAUDULEUSE OU NON À L'AIDE D'UNE FONCTION LOGISTIQUE.

LA FONCTION LOGISTIQUE

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

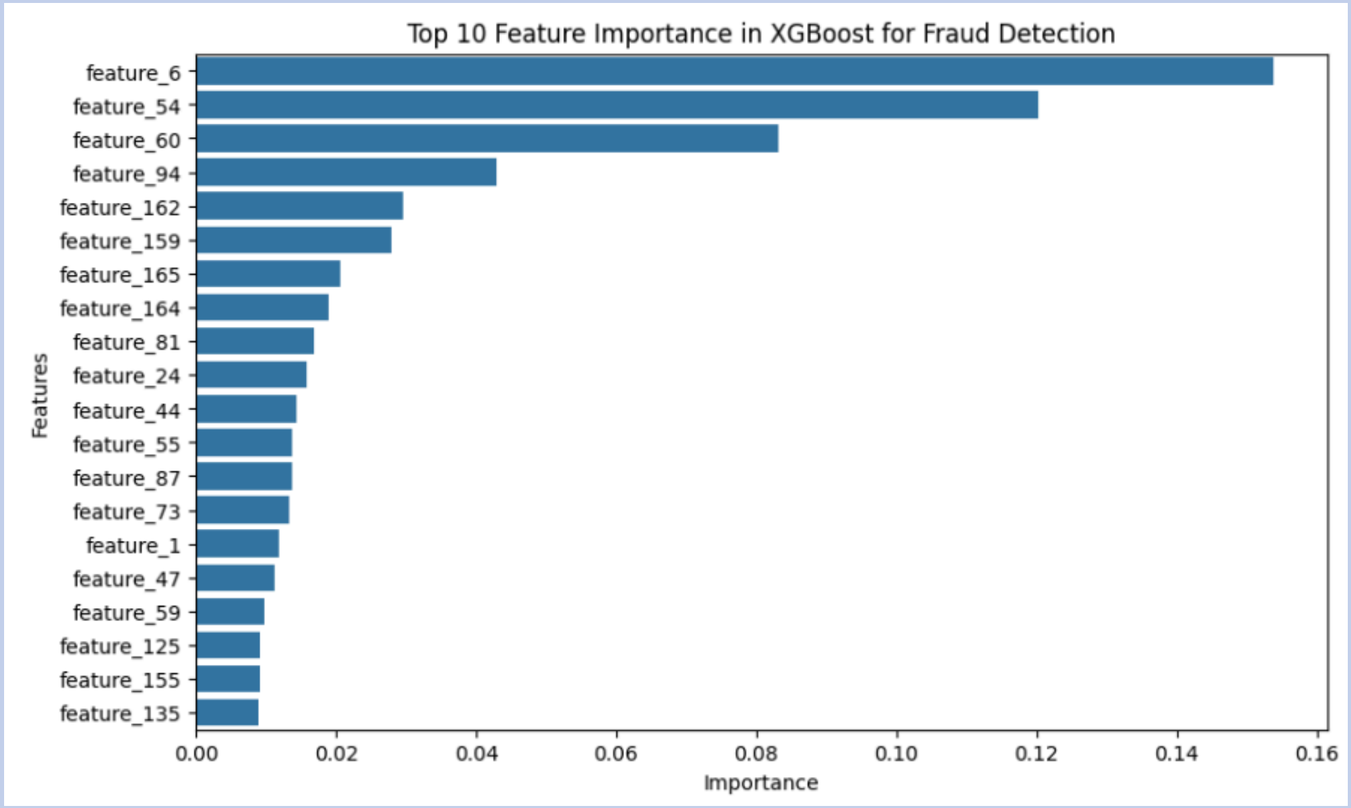




XGBOOST

XGBOOST EST UN ALGORITHME DE BOOSTING D'ARBRES DE DÉCISION QUI CONSTRUIT UNE SÉRIE DE MODÈLES FAIBLES ET LES COMBINE POUR AMÉLIORER LES PRÉDICTIONS.

OBJECTIF DANS NOTRE CAS : IDENTIFIER LES FEATURES LES PLUS IMPORTANTES POUR LA DÉTECTION DE FRAUDE, GRÂCE AUX SCORES D'IMPORTANCE CALCULÉS PAR L'ALGORITHME.

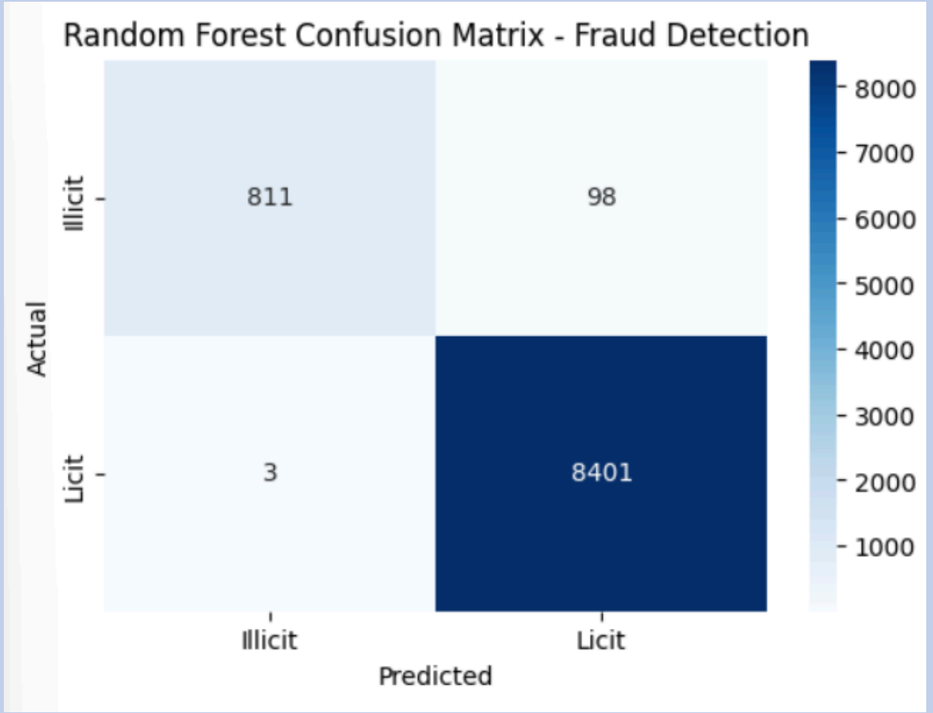


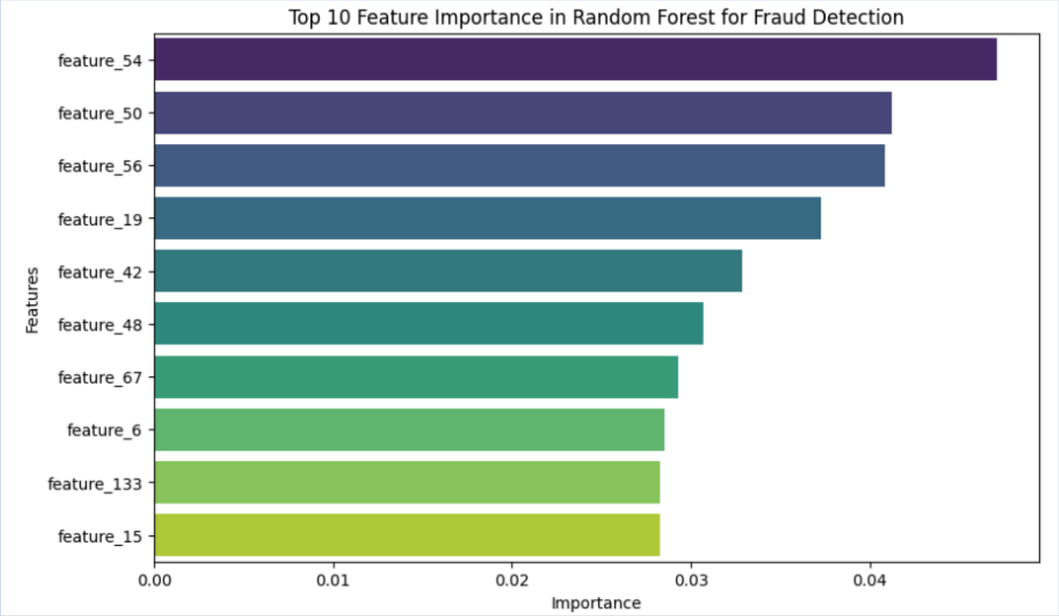
RANDOM FOREST

DÉFINITION SIMPLE : RANDOM FOREST EST UN ENSEMBLE D'ARBRES DE DÉCISION, OÙ CHAQUE ARBRE EST ENTRAÎNÉ SUR UN SOUS-ENSEMBLE DES DONNÉES ET DES FEATURES

OBJECTIF DANS NOTRE CAS :

- CLASSIFICATION BINAIRE
- SÉLECTION DES FEATURES IMPORTANTES





Modèle	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)
Logistic Regression	0.877	0.439	0.933	0.597
K-Nearest Neighbors	0.972	0.865	0.843	0.853
Decision Tree	0.981	0.887	0.917	0.902
Random Forest	0.989	0.996	0.892	0.941
ExtraTreesClassifier	0.988	0.996	0.879	0.934
XGBoost	0.991	0.995	0.912	0.952