PRÉSENTATION SUR L'UTILISATION DE TECHNIQUES D'APPRENTISSAGE AUTOMATIQUE POUR DÉTECTER LES FAUX BILLETS

BESOIN D'UN ALGORITHME ML SOPHISTIQUÉ

L'ONFCM a exprimé le besoin d'un algorithme de machine learning sophistiqué pour la détection rapide et efficace des faux billets.

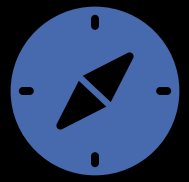
OBJECTIF PRINCIPAL

Développer un algorithme de ML performant pour déterminer l'authenticité des billets.

AVANTAGES DU ML

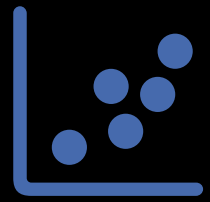
L'utilisation de méthodes avancées comme le ML permet une détection plus rapide et précise des faux billets.

Exploration **initiale** des données



COMPRENDRE LES DONNÉES

Explorer les données pour comprendre leur structure et leur contenu



EXPLORER LES RELATIONS

Explorer les relations entre les variables pour déterminer leurs effets sur le modèle

Axes du Notebook

COLLECTE DES DONNÉES

Des images de vrais et faux billets avec leurs étiquettes sont nécessaires.

PRÉTRAITEMENT

Les images subissent des transformations: niveaux de gris, redimensionnement, normalisation.

EXTRACTION DE CARACTÉRISTIQUES

Extraction de caractéristiques géométriques des billets: dimensions, longueurs, angles, proportions.

ENTRAÎNEMENT DU MODÈLE

Entraînement d'un classificateur SVM sur les données séparées en jeux d'entraînement et de test.

Données Géométrique

DÉTECTION DE FAUX BILLETS À L'AIDE DE LA…
COLLECTE DES DONNÉES

Le dataset contient 1500 en colonnes. Représentant 1000 billets authentiques ou 500 contrefaits

Données Géométrique

```python
# Afficher les informations sur le dataset
print("Information sur le dataset :")
print(billets.info())
```

✓ 0.0s

```
Information sur le dataset :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   is_genuine    1500 non-null   bool
 1   diagonal      1500 non-null   float64
 2   height_left   1500 non-null   float64
 3   height_right  1500 non-null   float64
 4   margin_low    1463 non-null   float64
 5   margin_up     1500 non-null   float64
 6   length        1500 non-null   float64
dtypes: bool(1), float64(6)
memory usage: 71.9 KB
None
```

Ces caractéristiques fournissent des informations importantes sur la forme et les dimensions physiques des billets, qui peuvent varier entre les vrais et faux.

En utilisant ces caractéristiques, le modèle peut identifier des billets, ce qui lui permet de faire des prédictions précises sur l'authenticité des billets inconnus.

Billets.describe

- La longueur diagonale des billets (diagonal) présente une moyenne de 171.96 et une valeur minimale de 171.04, avec un écart-type relativement faible de 0.31.

- Les hauteurs des côtés gauche (height_left) et droit (height_right) des billets ont des moyennes proches de 104, avec des écarts-types faibles, indiquant une certaine cohérence dans les dimensions.

- La marge inférieure des billets (margin_low) a été estimée en utilisant la moyenne pour les valeurs manquantes. La moyenne estimée est de 4.49, avec un écart-type de 0.66.

- La marge supérieure des billets (margin_up) présente une moyenne de 3.15 et un écart-type de 0.23.

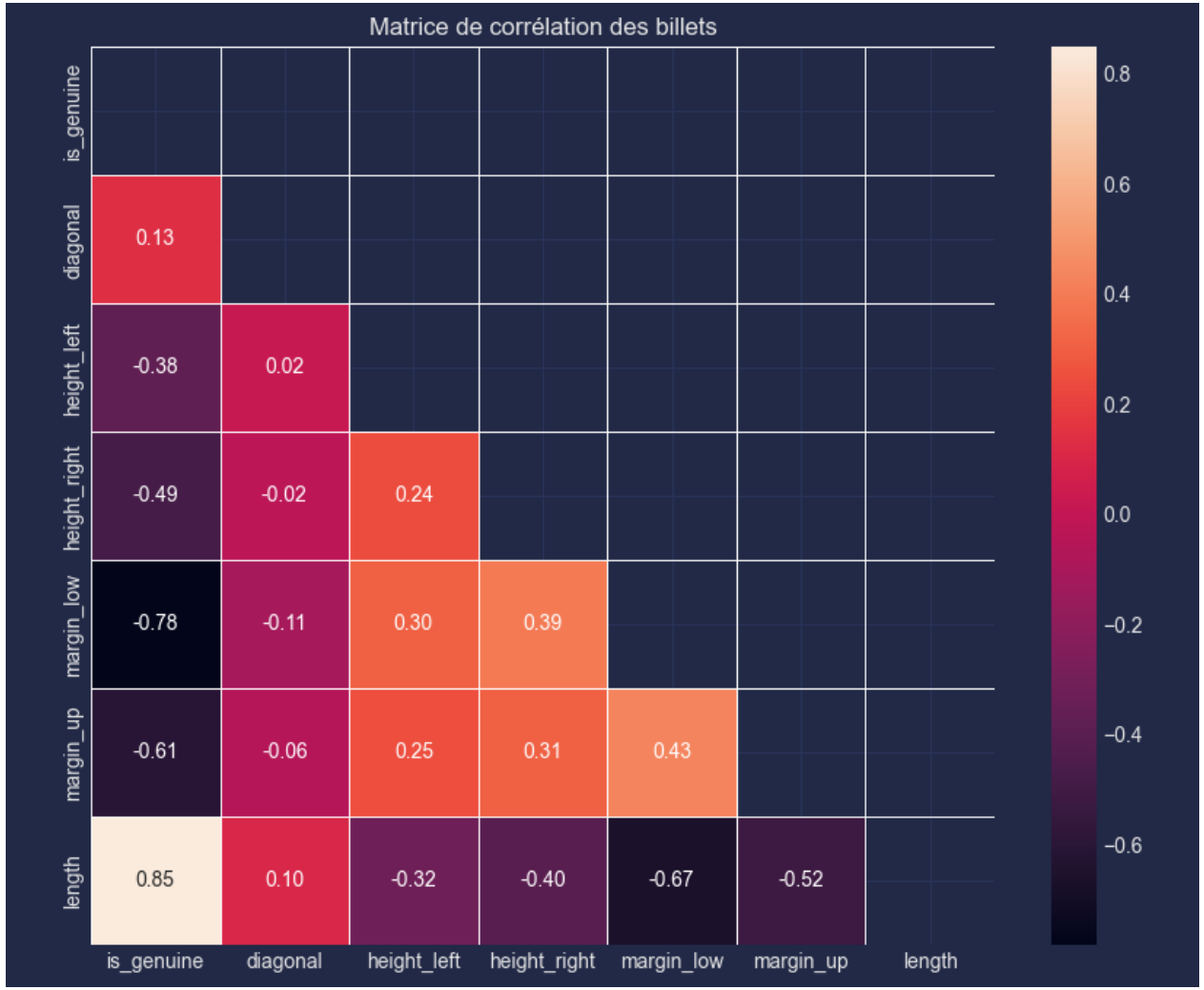
- La longueur totale des billets (length) a une moyenne de 112.68, avec un écart-type de 0.87.détection des faux billets.

Données Descriptives

Variable	Statistique
Diagonale	Moyenne: 171.96, Écart-type: 0.31
Hauteur G/D	Moyennes ~104, Écarts-types faibles
Marge basse	Moyenne: 4.49, Écart-type: 0.66
Marge haute	Moyenne: 3.15, Écart-type: 0.23
Longueur	Moyenne: 112.68, Écart-type: 0.87

*Résumé du dataset fourni

Matrice de corrélations

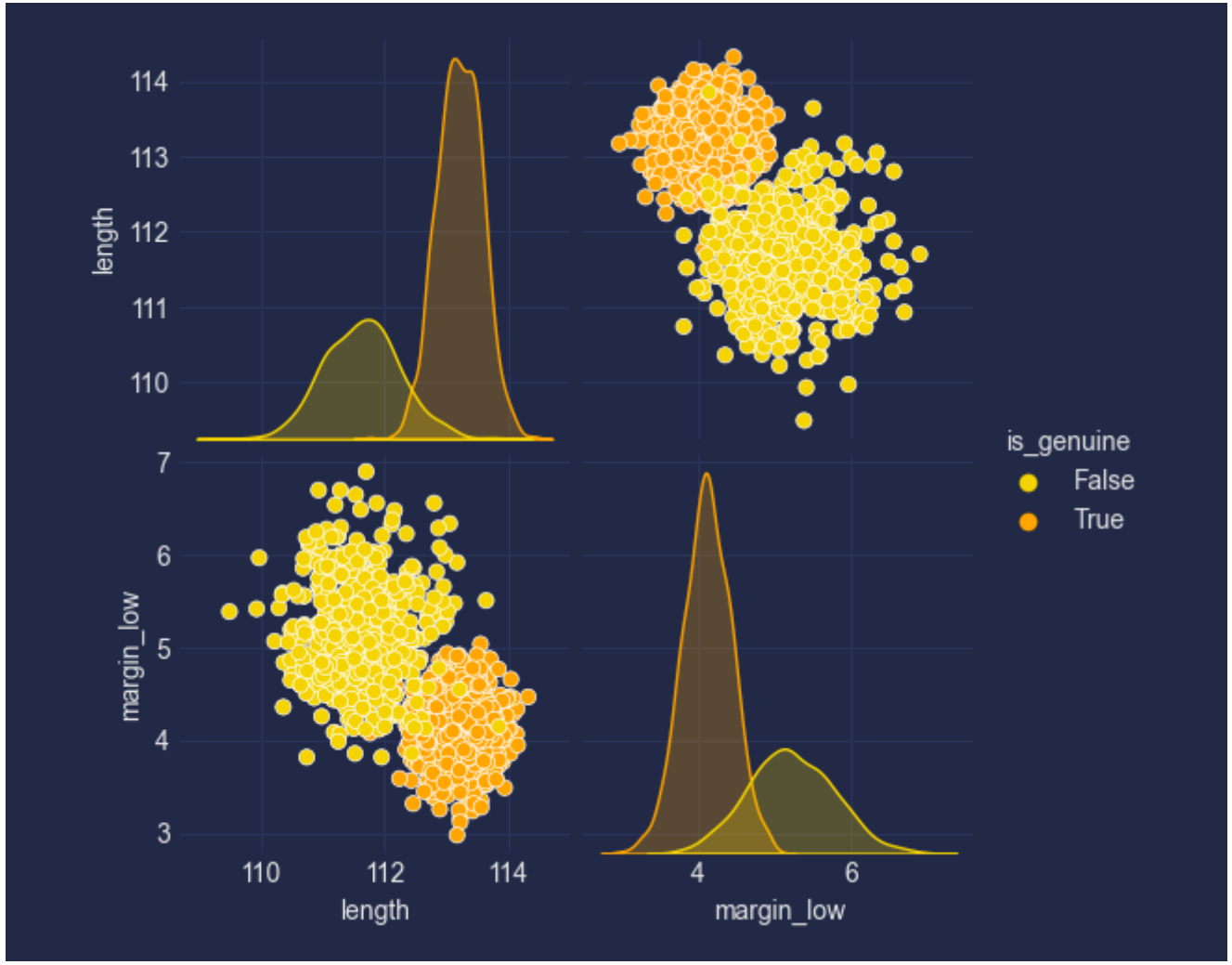


La distribution de la variable "diagonal" est similaire pour les vrais et les faux billets, ce qui suggère que cette variable ne sera pas ou peu utile dans notre prédiction.

En revanche, les variables "length" et "margin_low" semblent être des indicateurs pertinents, car nous pouvons clairement observer des différences dans leurs distributions entre les vrais et les faux billets.

En effet, lorsque nous examinons le nuage de points représentant la relation entre ces deux variables, nous pouvons distinguer deux groupes distincts.

Matrice de corrélations



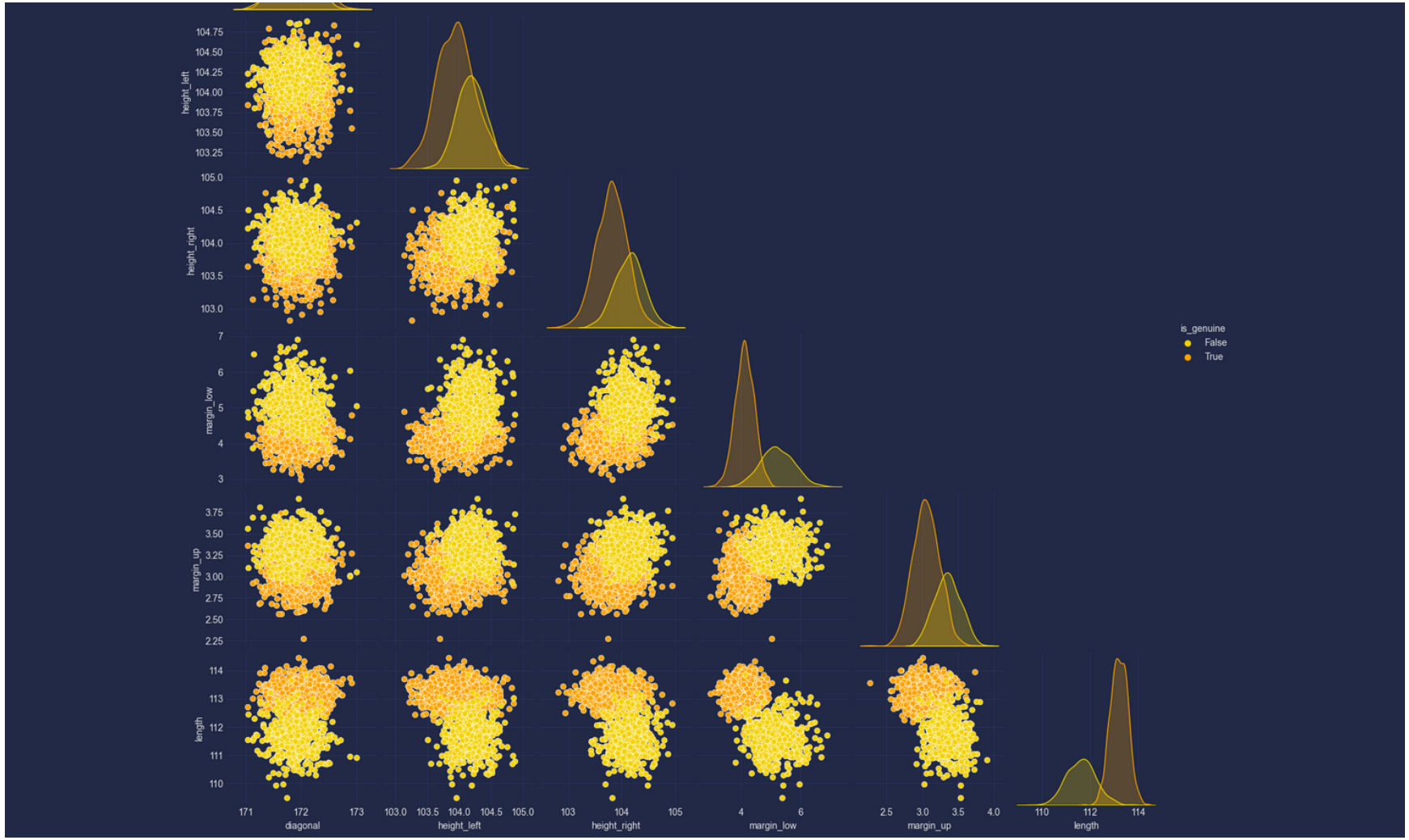
VISUALISATION DES RELATIONS ENTRE LES VARIABLES

En analysant le pairplot, nous recherchons des tendances, des motifs ou des corrélations entre les variables.

Certaines variables sont linéairement corrélées, des groupes distincts de valeurs se forment ou si des variables ont des distributions particulières.

C'est le cas Ici entre Length // Margin_low

Matrice de corrélations



VISUALISATION DES RELATIONS INTER VARIABLES

Nettoyage des données

DÉTECTION ET SUPPRESSION DES OUTLIERS

Détection Mais non suppression des valeurs car elle ne sont pas aberrantes d'apres une observation Box plot.

IMPUTATION DES DONNÉES MANQUANTES

37 Nan Margin_low

Imputation par regression linéaire .
Évaluation de la qualité de l'imputation.

STANDARDISATION DES DONNÉES

Standardisation par centrage-réduction pour éviter la domination de certaines variables.

Régression linéaire

RELATIONS SIGNIFICATIVES ENTRE VARIABLES INDÉPENDANTES ET DÉPENDANTE

Les coefficients de régression sont significatifs pour la plupart des variables indépendantes.

- Diagonal: -0.070

- Height_Left: 0.173

- Height_Right: 0.282

- Margin_Up: 0.203

- Length: -0.409

MODÈLE EXPLIQUE 48% DE LA VARIANCE

Le R2 de 0.482 indique que le modèle explique environ 48% de la variance de la variable dépendante

Le modèle semble avoir une performance modérée avec un R^2 de 0.474, indiquant que le modèle peut expliquer environ 47.4% de la variabilité dans les données.

ANALYSE DES RÉSIDUS NÉCESSAIRE

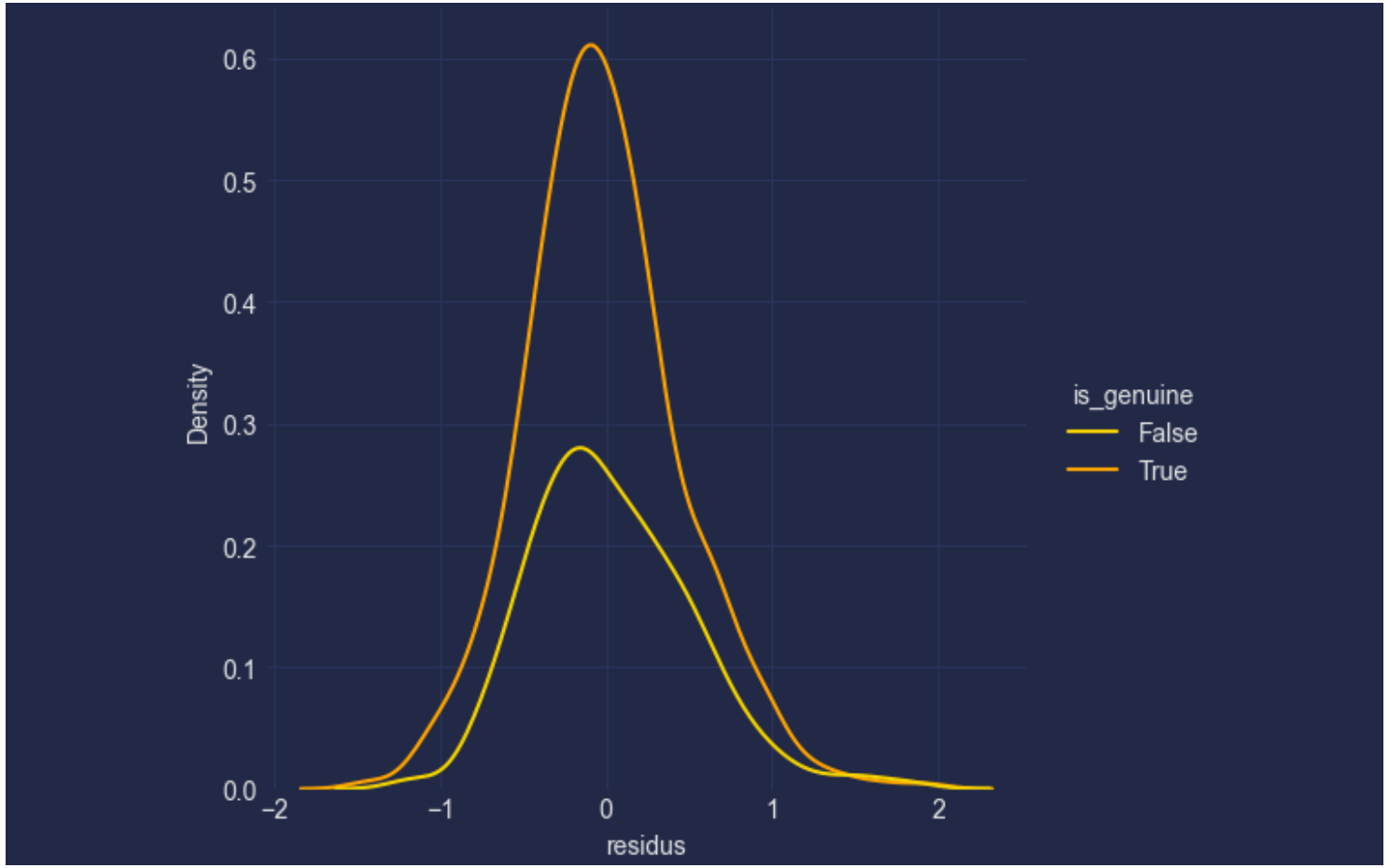
Les résidus du modèle, évalués par le test de Shapiro-Wilk, montrent une p-value très petite (1.17e-10), suggérant que les résidus ne suivent pas une distribution normale

Le coefficient pour "Diagonal" est -0.070 suggère une relation négative avec la variable dépendante "margin_low".

Pour chaque augmentation d'une unité de "Diagonal", la variable "margin_low" diminue, tout en gardant les autres variables constantes.

De même, le coefficient pour "Length" est -0.409, indiquant également une relation négative avec "margin_low".

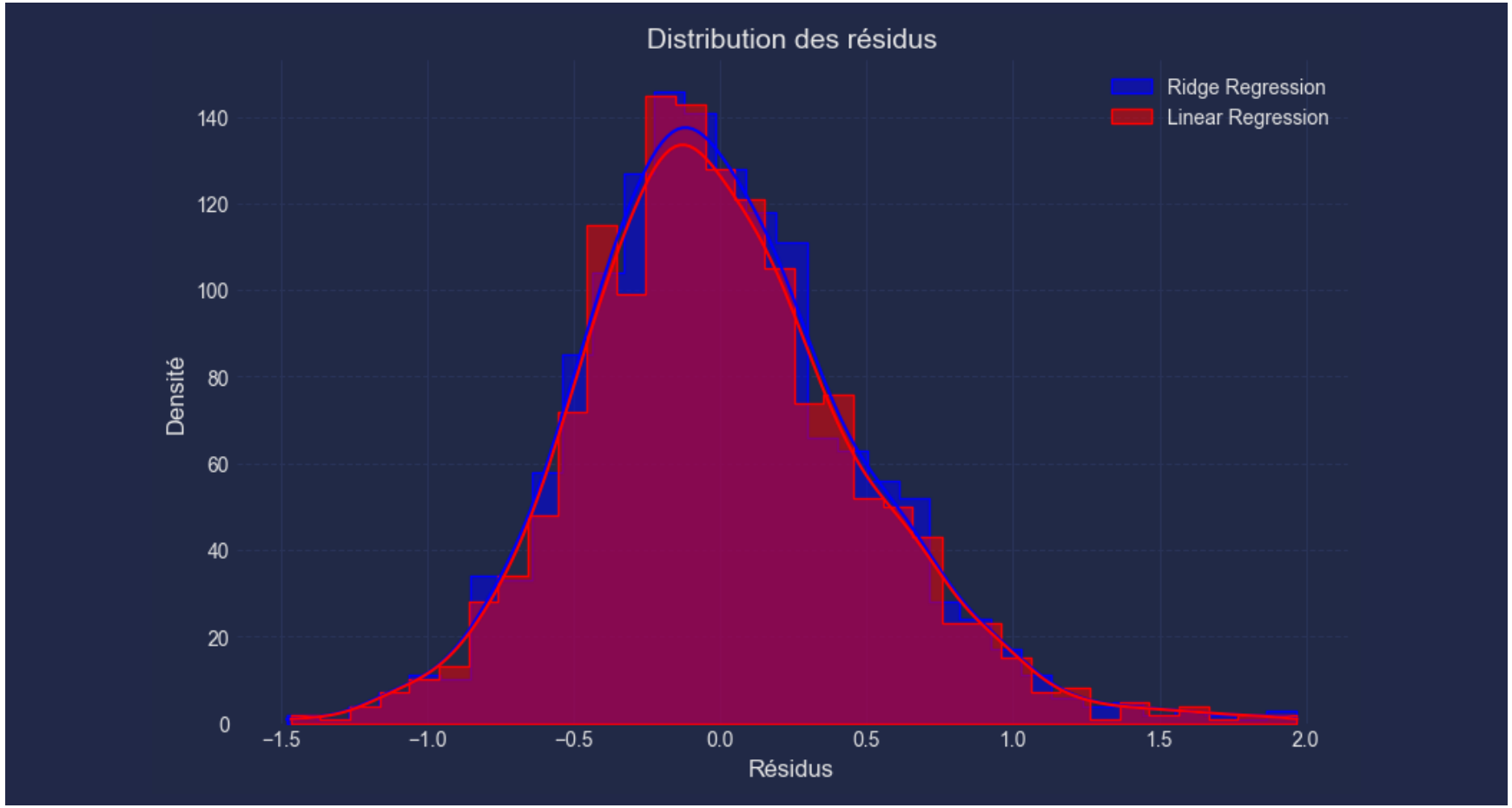
histogramme des probabilités prédites VS réels par le modèle de régression



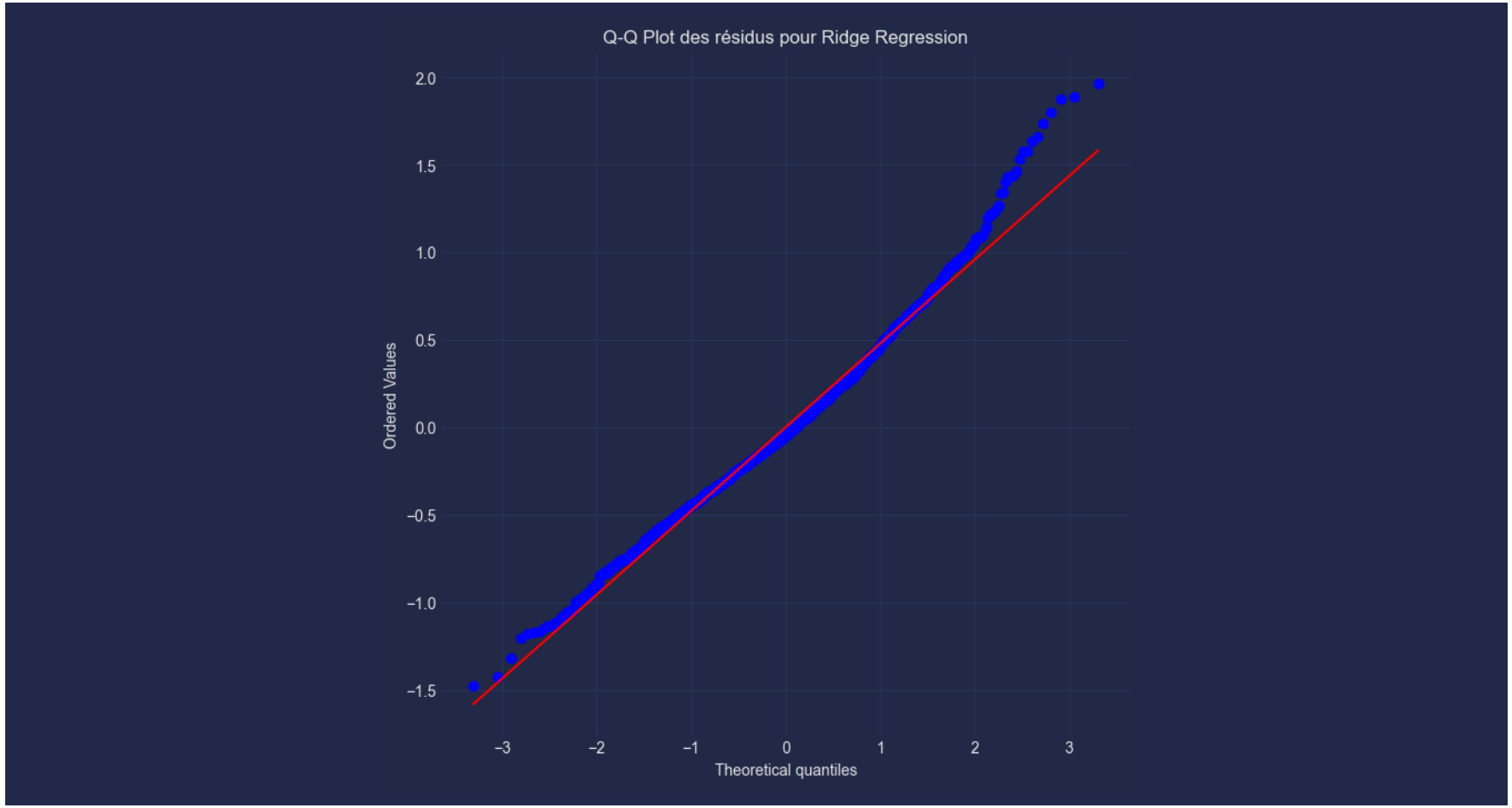
les résidus du modèle de régression logistique servent à vérifier la qualité de l'ajustement du modèle. Les résidus sont la différence entre les valeurs observées et les valeurs prédites par le modèle.

Un bon modèle de régression devrait avoir des résidus qui sont distribués de manière aléatoire et ne montrent pas de motif clair. ○

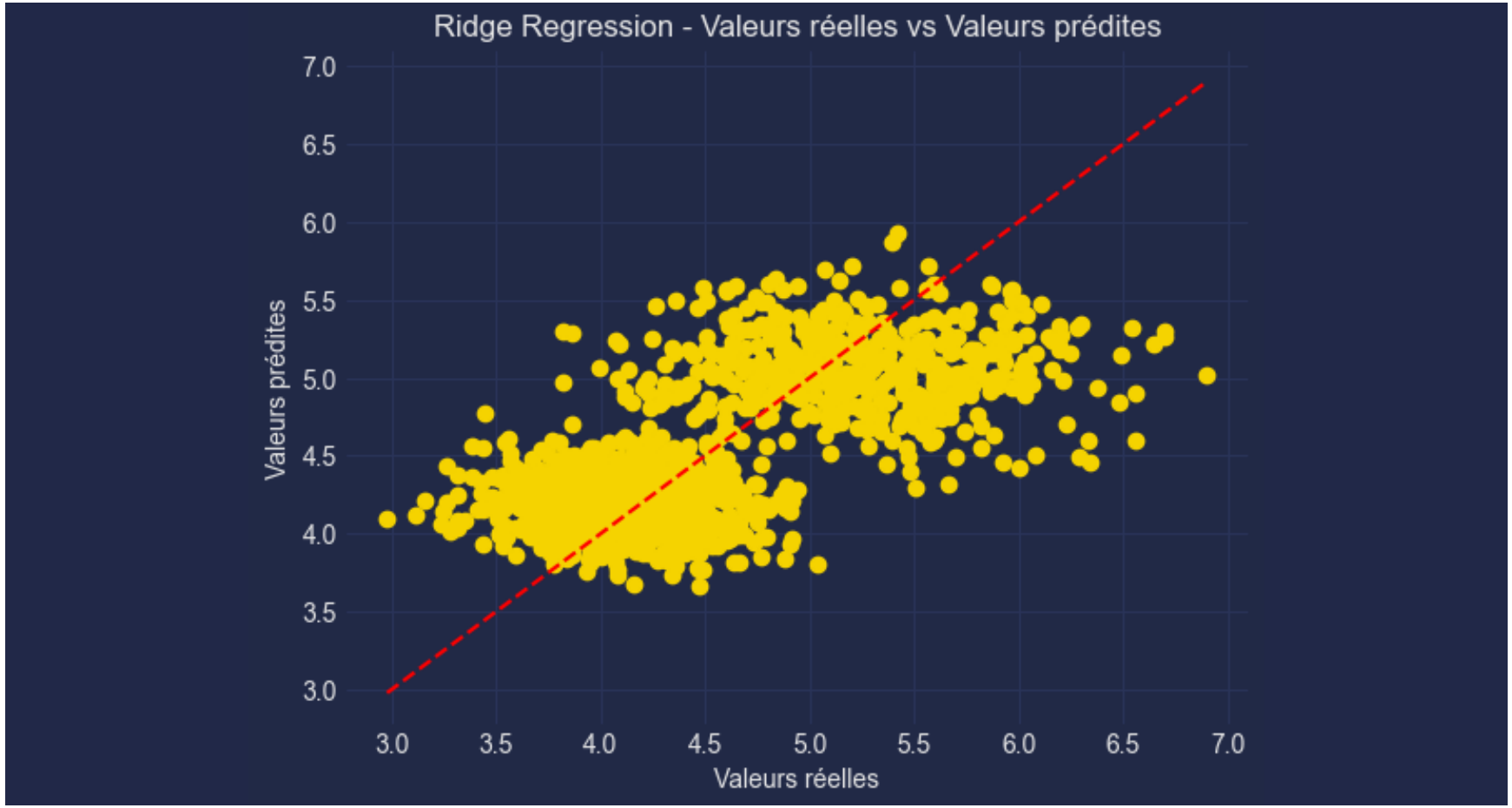
histogramme des probabilités prédites VS réels par le modèle de régression



histogramme des probabilités prédites VS réels par le modèle de régression



histogramme des probabilités prédites VS réels par le modèle de régression



Création de sous-ensembles de données

EXPLORATION DES DONNÉES

Explorer les données pour déterminer les sous-ensembles à créer

SÉLECTION DES CARACTÉRISTIQUES

Sélectionner les caractéristiques pertinentes pour la détection des faux billets

CRÉATION DES SOUS-ENSEMBLES

Créer des sous-ensembles de données à partir des caractéristiques sélectionnées

LA STANDARDISATION

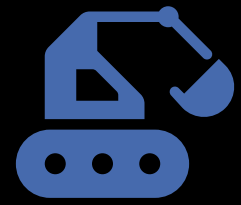
La standardisation peut être appliquée aux caractéristiques géométriques des billets utilisées pour entraîner les modèles de détection.

Ces caractéristiques peuvent inclure la longueur, la largeur, la hauteur et d'autres dimensions du billet.

En standardisant ces caractéristiques, on s'assure que toutes les dimensions sont mises à la même échelle, ce qui facilite la comparaison et l'analyse des données.

Centrage des données :

La première étape de la standardisation consiste à soustraire la moyenne de chaque variable, de sorte que la moyenne de la variable transformée soit égale à zéro,Cela permet de centrer les données autour de zéro.

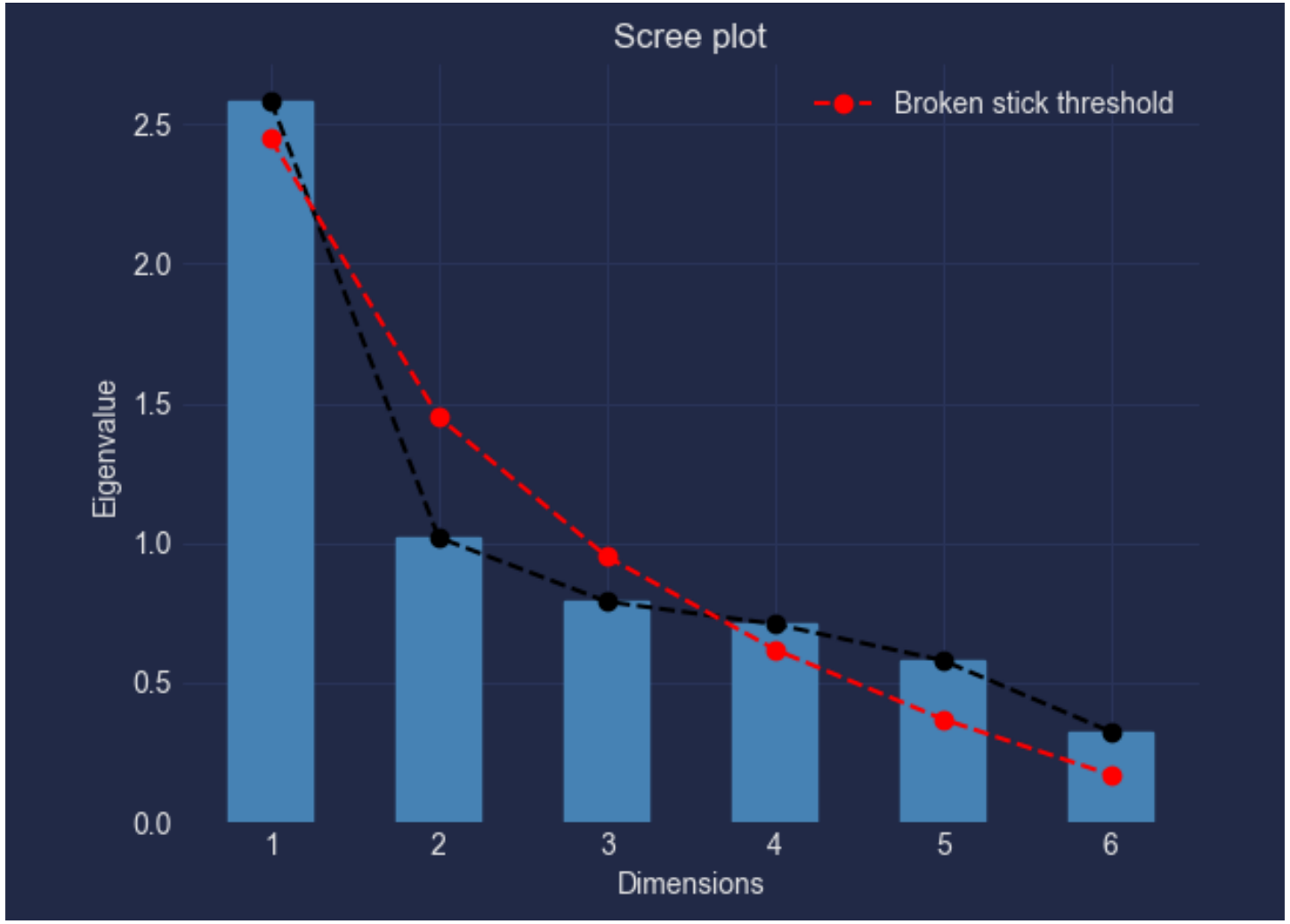


PRINCIPES DE BASE DE LA STANDARDISATION

Réduction des écarts : Ensuite, on divise chaque valeur de la variable centrée par l'écart type de la variable d'origine. Cela permet de réduire les écarts entre les valeurs et d'obtenir une variance unitaire.

Mise à l'échelle : La standardisation met toutes les variables sur une même échelle, ce qui signifie qu'elles ont toutes une variance égale à un.

ACP



ANALYSE EN COMPOSANTES PRINCIPALES (ACP) :

LES VARIABLES EXPLICATIVE

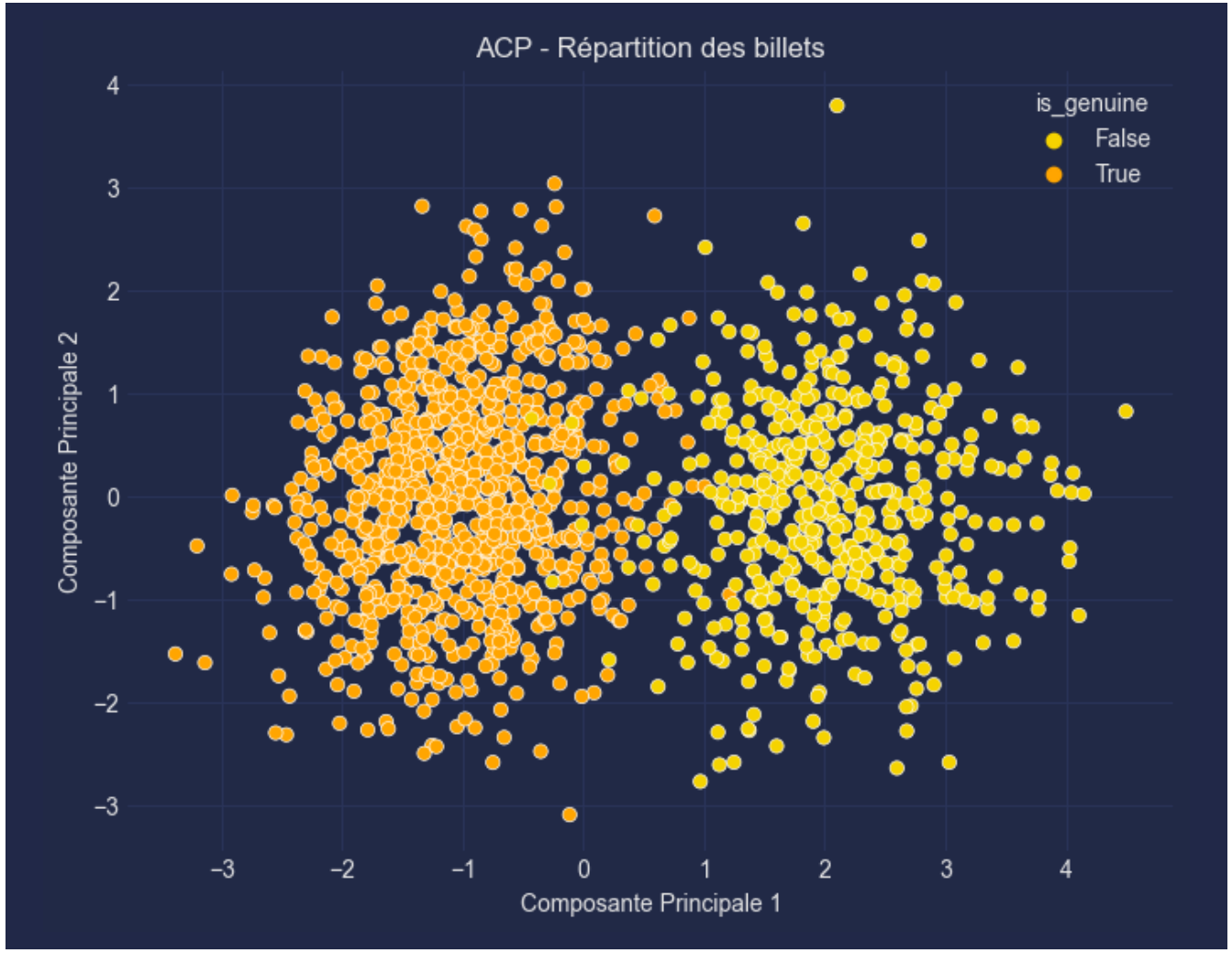
SONT STANDARDISÉES

AVANT DE RÉALISER L'ACP,

RÉDUISANT LES DONNÉES

À DEUX COMPOSANTES PRINCIPALES.

ACP



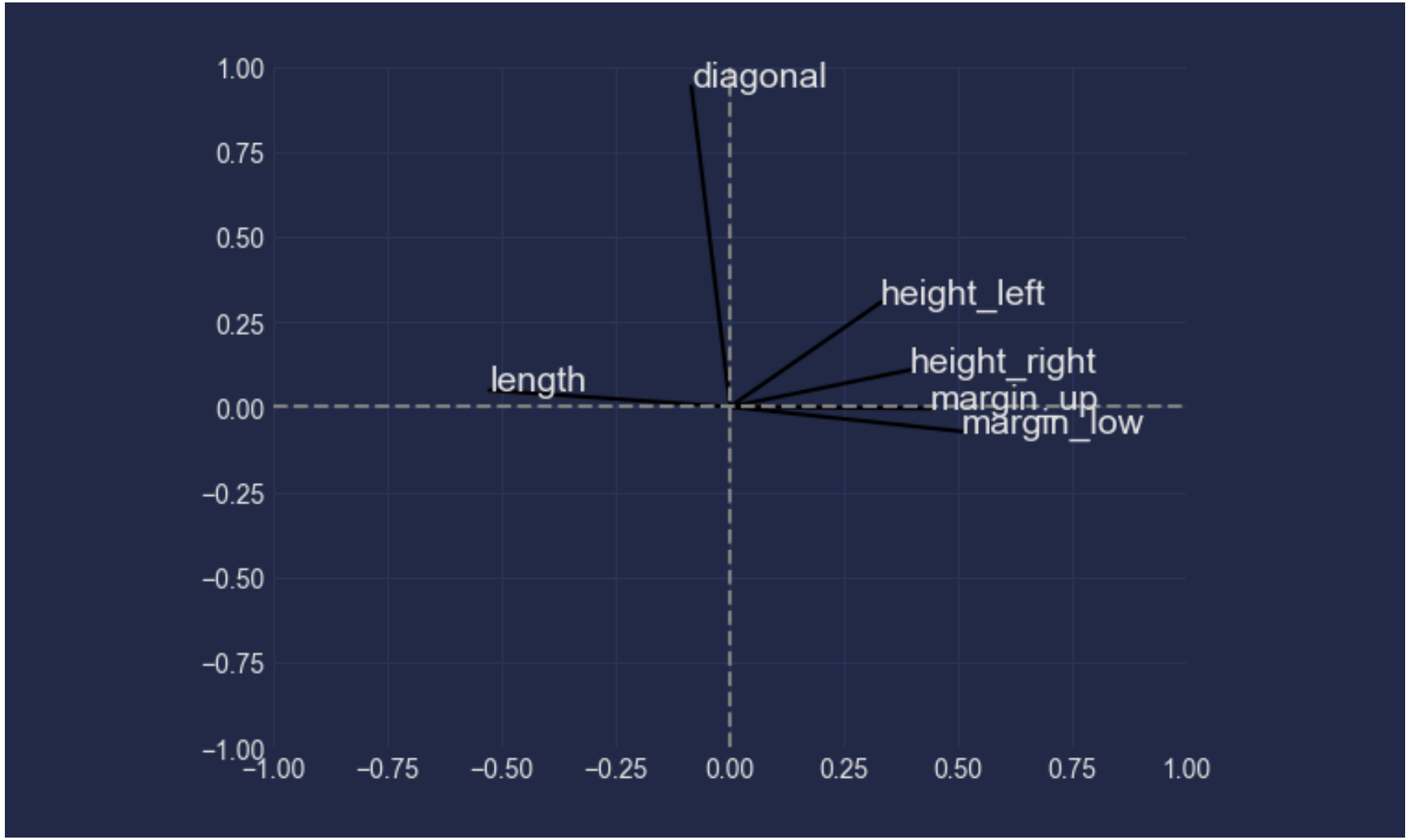
VISUALISATION DES RÉSULTATS DE L'ACP

APRÈS AVOIR PROJETÉ NOS DONNÉESEN COLORANT LES POINTS

EN FONCTION DE LEUR ÉTIQUETTE DE VÉRACITÉ (VRAI OU FAUX BILLETS)

NOUS POUVONS CLAIREMENT VOIR LA DISTINCTION ENTRE LES DEUX GROUPES.

ACP



CETTE VISUALISATION CONFIRME NOS OBSERVATIONS PRÉCÉDENTES SELON LESQUELLES CERTAINES VARIABLES,

TELLES QUE LA LONGUEUR (LENGTH) ET LA MARGE BASSE (MARGIN_LOW),

JOUENT UN RÔLE IMPORTANT DANS LA PRÉDICTION DE LA VÉRACITÉ DES BILLETS.

LES DEUX GROUPES DISTINCTS SUGGÈRENT QU'IL EXISTE DES CARACTÉRISTIQUES GÉOMÉTRIQUES SPÉCIFIQUES QUI DIFFÉRENCIENT LES VRAIS BILLETS DES FAUX BILLETS.

K-means

PERFORMANCE DE K-MEANS

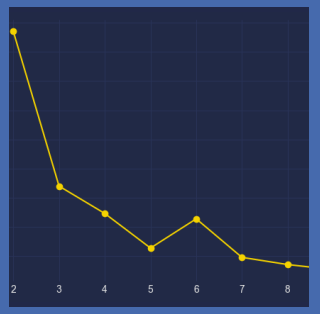
K-means a montré une performance remarquable pour la classification des billets authentiques et contrefaits

Accuracy: 0.9813333333333333

Precision: 0.9869739478957916

Recall: 0.985

F1-score: 0.9859859859859861



NOMBRE OPTIMAL DE CLUSTERS

Le choix de deux clusters s'est avéré efficace pour séparer les observations en authentiques et contrefaits

La moyenne silhouette_score est : 0.34274600198642563

MÉTRIQUES DE VALIDATION

Le score de silhouette et le score de Davies-Bouldin ont validé le choix de deux clusters

Le score de Davies-Bouldin est : 1.2153573863170573

SVM

PRÉCISION EXCEPTIONNELLE

Le modèle SVM a démontré une précision de 0.9911 dans la classification des billets authentiques et contrefaits.

ROBUSTESSE

Le modèle SVM s'est avéré très robuste, performant de manière stable même avec de nouvelles données.

MEILLEURE ACCURACY

L'accuracy de 0.992 obtenue par le modèle SVM surpasse celle des autres modèles évalués.

MEILLEUR F1-SCORE

Avec un F1-score de 0.9940, le modèle SVM a obtenu le meilleur score de F1 parmi tous les modèles.

MODÈLE DE CHOIX

Compte tenu de sa précision et de sa robustesse exceptionnelles, le modèle SVM est le modèle de choix pour la détection des faux billets.

PERFORMANCE GLOBALE

Le modèle SVM s'est distingué par sa performance globale, surpassant les autres modèles sur les métriques clés d'évaluation.

Résultats Rappel

RÉGRESSION LINÉAIRE

RMSE de 0.481 et R2 de 0.474 indiquent une performance modérée pour prédire les valeurs.

RÉGRESSION LOGISTIQUE

Précision de 0.9967 est extrêmement élevée pour classifier les billets vrais/faux.

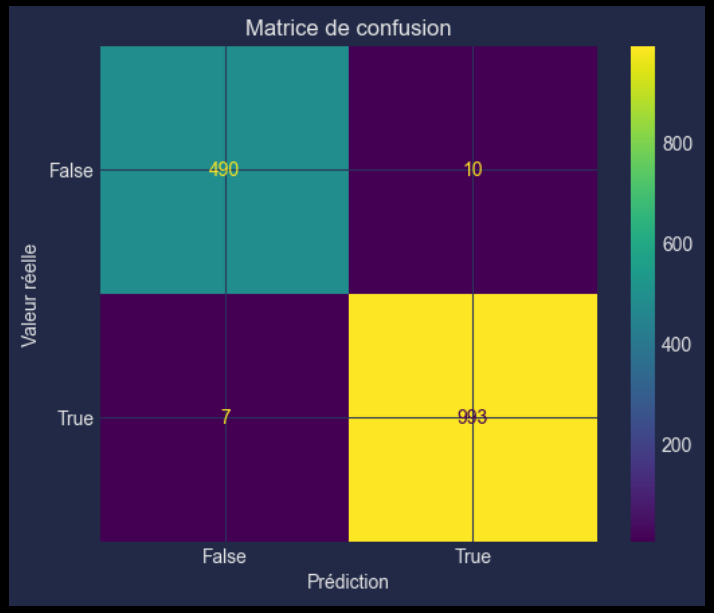
K-MEANS

Précision de 0,98 la méthode à bien séparé les groupes.

EN RÉSUMÉ

La SVM est la plus efficace. La régression linéaire est modérément bonne. Le K-means a de plus faibles performances.

Résultats



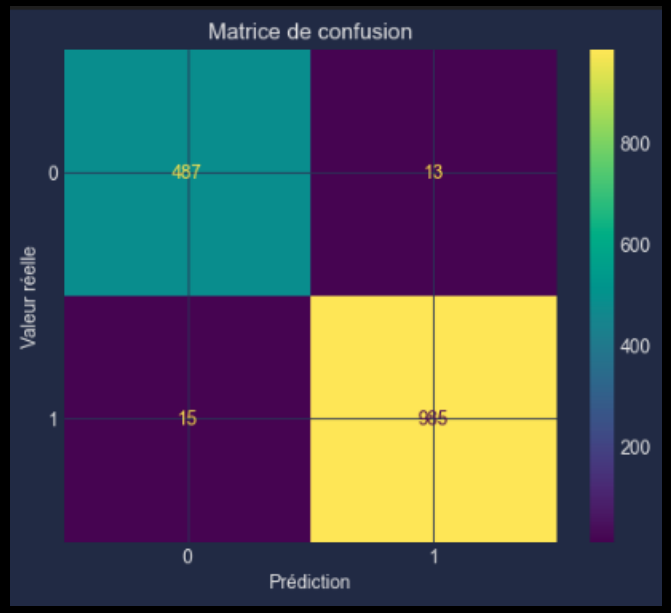
RÉSULTATS DES TESTS RÉGRESSION LOGISTIQUES

Accuracy: 0.9886666666666667

Precision: 0.9900299102691924

Recall: 0.993

F1-score: 0.9915127309036444



RÉSULTATS DES TESTS K-MEANS

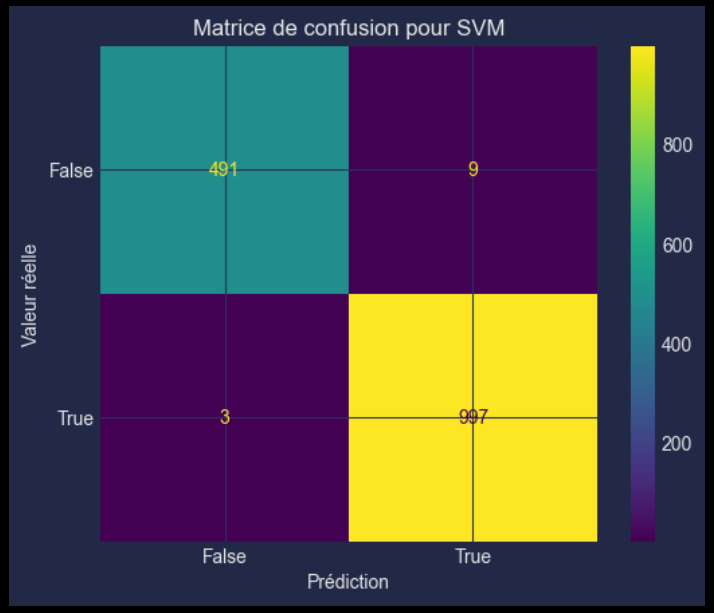
Accuracy: 0.9813333333333333

Precision: 0.9869739478957916

Recall: 0.985

F1-score: 0.9859859859859861

Le score de Davies-Bouldin est : 1.2153573863170573



RÉSULTATS DES TESTS SVM

SVM Accuracy: 0.992

SVM Precision: 0.9910536779324056 SVM Recall: 0.997

SVM F1-score: 0.9940179461615155

ALGO_SVM

PERFORMANCE DE SVM

Le modèle SVM se distingue par sa robustesse et sa précision exceptionnelle, avec un F1-score de 0.9940

LE MODÈLE SVM EST LE PLUS PERFORMANT POUR LA DÉTECTION DES FAUX BILLETS.