# Analyse comparative des modèles de régression logistique et K-means

Aurélien Bleusez

22 février 2024

**Abstract**

L'objectif de ce projet est de développer une application en Python de machine learning pour l'identification des faux billets. Nous comparons les performances de deux approches : un modèle de classification supervisée utilisant la régression logistique et un modèle non supervisé de clustering avec l'algorithme K-Means. Notre jeu de données comprend les caractéristiques de 1500 billets, telles que la longueur, la largeur, la hauteur, et d'autres mesures pertinentes. Le développement de ce système vise à fournir un outil efficace pour lutter contre la fraude financière.

Les performances des modèles supervisés et non supervisés sont similaires dans notre étude. Cependant, dans le cadre d'une application de détection de faux billets, un modèle de régression logistique semble plus approprié pour plusieurs raisons :

1. Interprétabilité : La régression logistique fournit une interprétation directe des coefficients, permettant de comprendre l'importance de chaque variable dans la prédiction, ce qui est crucial dans le contexte de la détection de faux billets.
2. Facilité d'ajustement : Les modèles de régression logistique sont relativement simples à ajuster et à interpréter. Avec des paramètres bien choisis et un prétraitement approprié des données, il est souvent possible d'obtenir des performances élevées avec ce type de modèle.

En conclusion, bien que les performances des modèles supervisés et non supervisés soient comparables, un modèle de régression logistique offre des avantages significatifs en termes d'interprétabilité et de facilité d'ajustement, ce qui en fait un choix plus approprié pour une application de détection de faux billets.

# Sommaire

# 1 Introduction

La lutte contre la fraude financière est une préoccupation majeure dans de nombreux secteurs. Dans ce contexte, l'identification des faux billets est une tâche cruciale pour les institutions financières, les commerces de détail et les organismes chargés de l'application de la loi. L'objectif de ce projet est de développer une application en Python de machine learning capable de détecter les faux billets.

Nous nous concentrons sur la comparaison des performances de deux approches : un modèle de classification supervisée utilisant la régression logistique et un modèle non supervisé de clustering avec l'algorithme K-Means. Pour ce faire, nous utilisons un jeu de données comprenant les caractéristiques de 1500 billets, telles que la longueur, la largeur, la hauteur, et d'autres mesures pertinentes.

Le développement de ce système vise à fournir un outil efficace pour aider les institutions à lutter contre la fraude financière en identifiant rapidement et avec précision les faux billets.

# 2 Méthode

Dans cette section, nous décrivons les méthodes et les outils utilisés pour développer notre application de détection de faux billets.
Nous avons principalement utilisé les bibliothèques suivantes :

- **Python :** Langage de programmation principal pour le développement de l'application.
- **Scikit-learn :** Bibliothèque Python pour l'apprentissage automatique, utilisée pour mettre en œuvre les modèles de classification supervisée et de clustering non supervisé.
- **NumPy :** Bibliothèque Python pour le calcul numérique, utilisée pour la manipulation efficace des données.

Le livrable final de ce projet est une application en ligne de commande, offrant une interface simple pour l'utilisation des modèles de détection de faux billets.

# 3 Exploration

L'exploration des données permet de comprendre la structure du jeu de données, ainsi que d'identifier les incohérences et les valeurs manquantes. Durant cette phase, aucune valeur aberrante n'a été détectée dans le jeu de données. Cependant, il est important de noter la présence de 37 valeurs manquantes dans la colonne `margin_low`, ce qui représente environ 2.5% des données (table 2). L'affichage des donées sous forme matricielle permet d'apréhender leurs distributions (figure 1).

Table 1: Répartition des données

| is_genuine | count |
|---|---|
| False | 500 |
| True | 1000 |

Table 2: Résumé des statistiques des données

| | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|
| **count** | 1500 | 1500 | 1500 | 1463 | 1500 | 1500 |
| **nan_count** | 0 | 0 | 0 | 37 | 0 | 0 |
| **nan_ratio %** | 0 | 0 | 0 | 2.47 | 0 | 0 |
| **mean** | 171.96 | 104.03 | 103.92 | 4.49 | 3.15 | 112.68 |
| **std** | 0.31 | 0.3 | 0.33 | 0.66 | 0.23 | 0.87 |
| **min** | 171.04 | 103.14 | 102.82 | 2.98 | 2.27 | 109.49 |
| **25%** | 171.75 | 103.82 | 103.71 | 4.01 | 2.99 | 112.03 |
| **50%** | 171.96 | 104.04 | 103.92 | 4.31 | 3.14 | 112.96 |
| **75%** | 172.17 | 104.23 | 104.15 | 4.87 | 3.31 | 113.34 |
| **max** | 173.01 | 104.88 | 104.95 | 6.9 | 3.91 | 114.44 |

**Scatter Matrix**

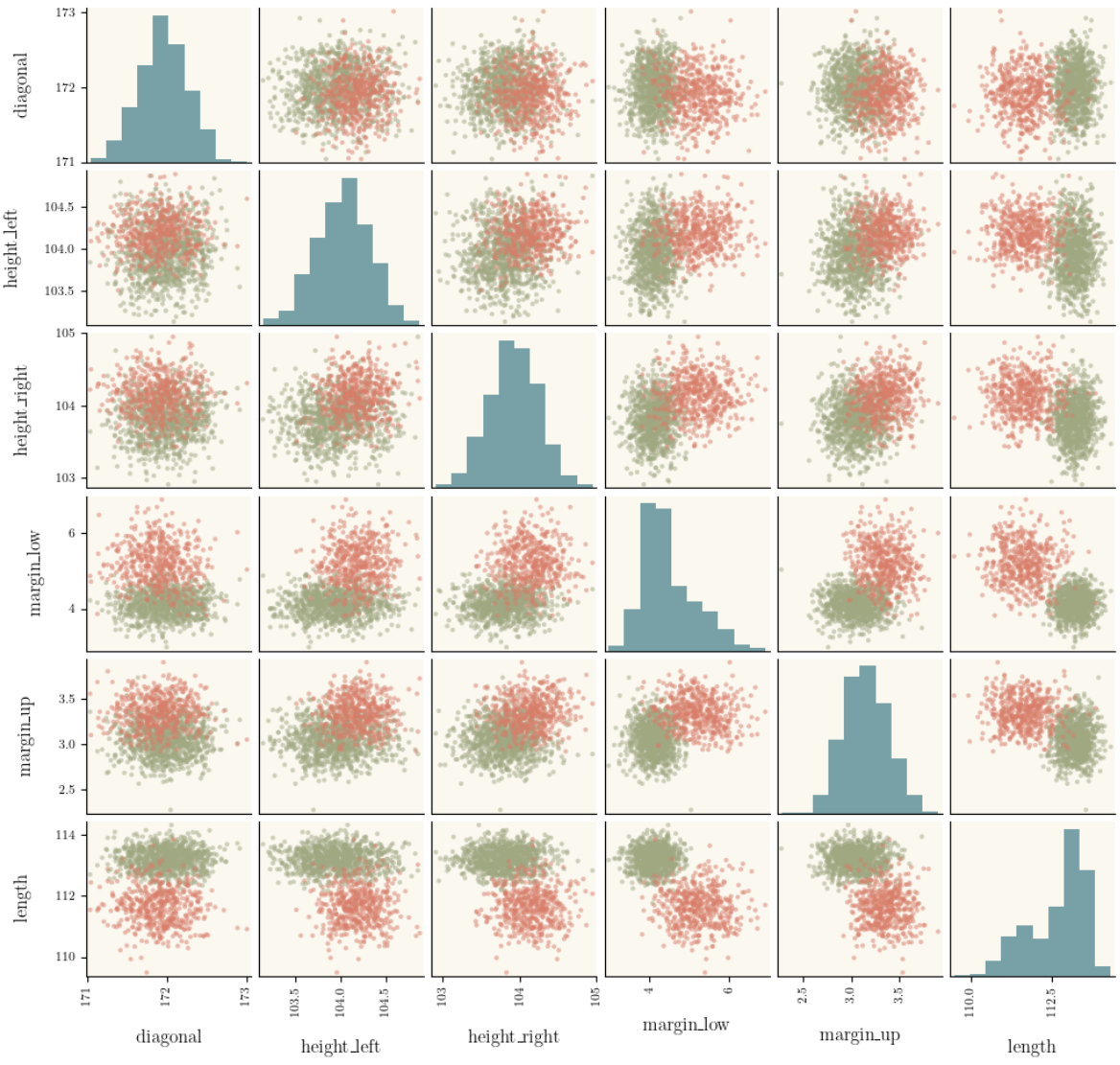



Figure 1: Data Visualization

Dans l'ensemble, les données suivent une distribution gausséenne aussi à ce stade nous prenons la décision de ne pas appliquer d'autres transformation que la standardisation des données. On observe une nette distinction des masses de vrais et faux billet sur les figures des variables `margin_low`, `margin_up`, `length` . Nous supposons que ces variables seront les coefficients les plus importants de notre modèle de régression logistique. L'affichage des box plots (figure 2) met en avant les différences et ressemblances entre les catégories de billets et permet une confirmation visuelle de l'importance de ces variables.

Les données étant des mesures géométriques, il est raisonnable de s'attendre à des phénomènes de multicolinéarité. Afin d'évaluer l'ampleur de cet effet, nous avons calculé le facteur d'inflation de la variance (VIF) pour chaque variable (table 3). Les valeurs calculées indiquent la présence de colinéarités dont l'impact semble être négligeable, avec des VIFs inférieurs à 5. À ce stade, nous décidons de conserver l'ensemble des variables et de procéder à une sélection des variables lors de l'évaluation des méthodes de prétraitement

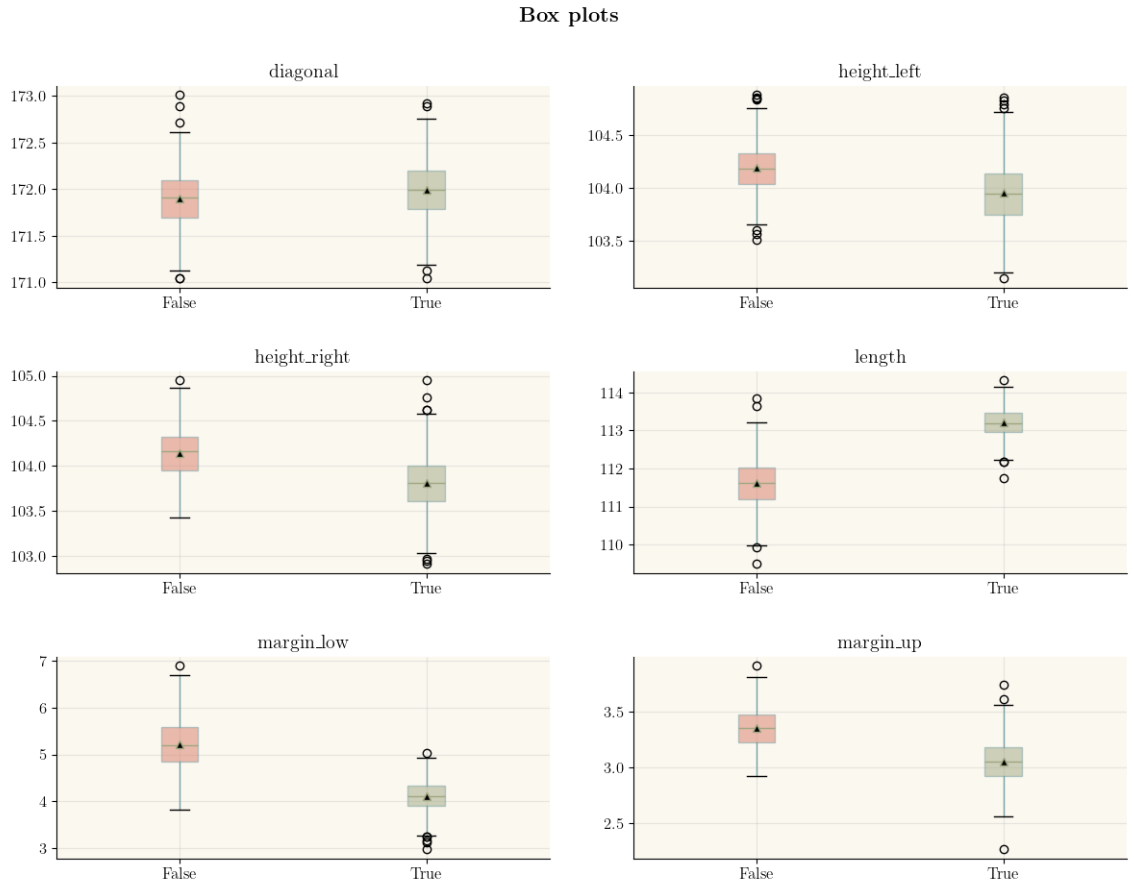



Figure 2: boxplot

Table 3: Facteur d'inflation de la variance (VIF)

| **Variable** | **VIF** |
|---|---|
| diagonal | 1.01861 |
| height_left | 1.15147 |
| height_right | 1.26029 |
| margin_low | 1.91328 |
| margin_up | 1.41967 |
| length | 2.13107 |

# 4 Prétraitement

Dans le cadre de ce projet, nous avons mis en œuvre une étape de prétraitement pour gérer les données manquantes. Afin de pallier ces valeurs manquantes, nous avons exploré trois méthodes d'imputation : la suppression des lignes concernées, l'imputation par régression linéaire, et l'imputation par la méthode des k plus proches voisins (KNN) (cf. table 4 & 5)

## 4.1 Régression Logistique

Table 4: Cross validation des méthodes d'imputation. Kfold splits : 3

| | **fit_time** | **score_time** | **log_loss_score** | **roc_auc_score** |
|---|---|---|---|---|
| **Drop NaN** | | | | |
| **mean** | 0.002712 | 0.002579 | -0.036859 | 0.998437 |
| **min** | 0.002486 | 0.002447 | -0.041870 | 0.996293 |
| **max** | 0.003046 | 0.002723 | -0.032890 | 0.999887 |
| **Régression linéaire** | | | | |
| **mean** | 0.002475 | 0.002424 | -0.036405 | 0.998242 |
| **min** | 0.002409 | 0.002333 | -0.047998 | 0.996476 |
| **max** | 0.002573 | 0.002522 | -0.022035 | 0.999928 |
| **KNN imputer** | | | | |
| **mean** | 0.003873 | 0.003754 | -0.037545 | 0.998194 |
| **min** | 0.003763 | 0.003742 | -0.048402 | 0.996278 |
| **max** | 0.003930 | 0.003777 | -0.022679 | 0.999928 |

## 4.2 K-means

Table 5: Cross validation des méthodes d'imputation. Kfold splits : 3

| | **fit_time** | **score_time** | **v_measure** | **rand_index** |
|---|---|---|---|---|
| **Drop NaN** | | | | |
| **mean** | 0.013064 | 0.002886 | 0.895551 | 0.972987 |
| **min** | 0.002046 | 0.002629 | 0.887137 | 0.967684 |
| **max** | 0.034955 | 0.003232 | 0.901485 | 0.975662 |
| **Régression linéaire** | | | | |
| **mean** | 0.002120 | 0.002668 | 0.900992 | 0.974961 |
| **min** | 0.001969 | 0.002632 | 0.879595 | 0.968449 |
| **max** | 0.002219 | 0.002720 | 0.935993 | 0.984096 |
| **KNN** | | | | |
| **mean** | 0.031517 | 0.004241 | 0.896277 | 0.973665 |
| **min** | 0.002088 | 0.002645 | 0.873245 | 0.968449 |
| **max** | 0.083317 | 0.007422 | 0.935993 | 0.984096 |

## Analyse

Étant donné la similarité des résultats obtenus, nous aurions pu choisir de supprimer les données manquantes, comme le confirme la courbe d'apprentissage démontrant une quantité de données suffisante (cf. figure 3). Cependant, dans le cadre de cette étude, nous avons opté pour l'imputation par régression linéaire. Les performances de cette méthode sont corroborées visuellement par l'analyse de son homoscédasticité (cf. figure 3.1).

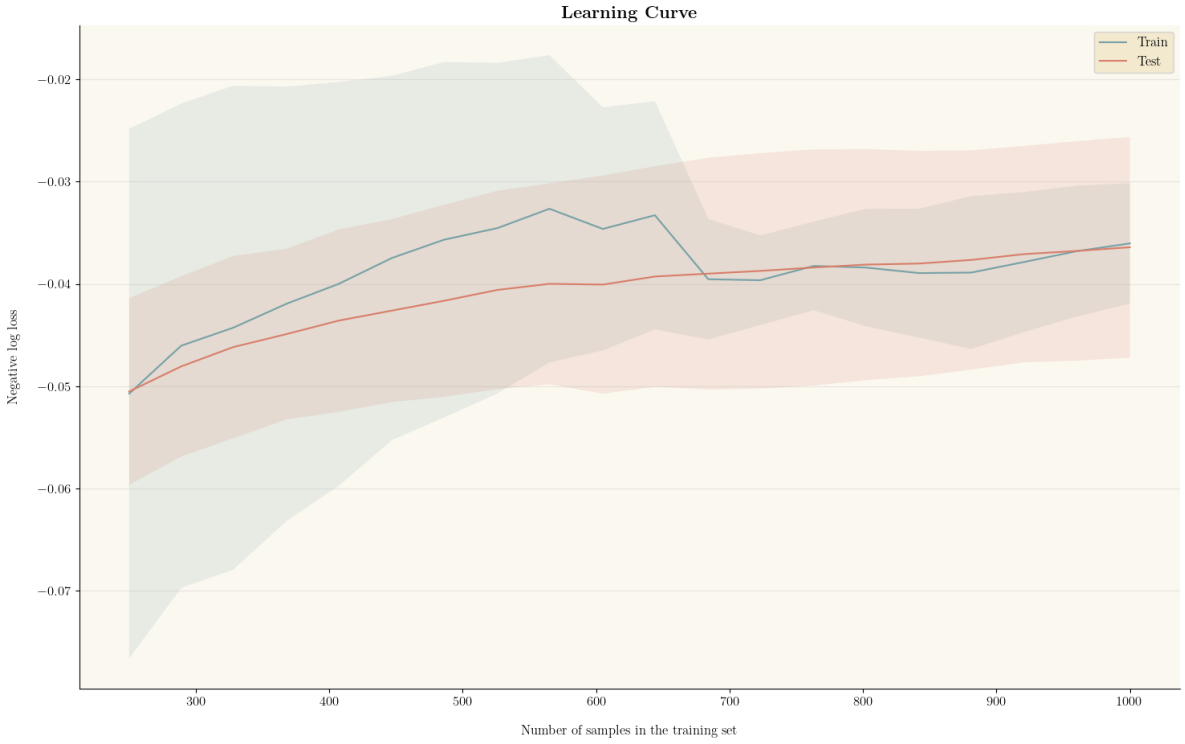
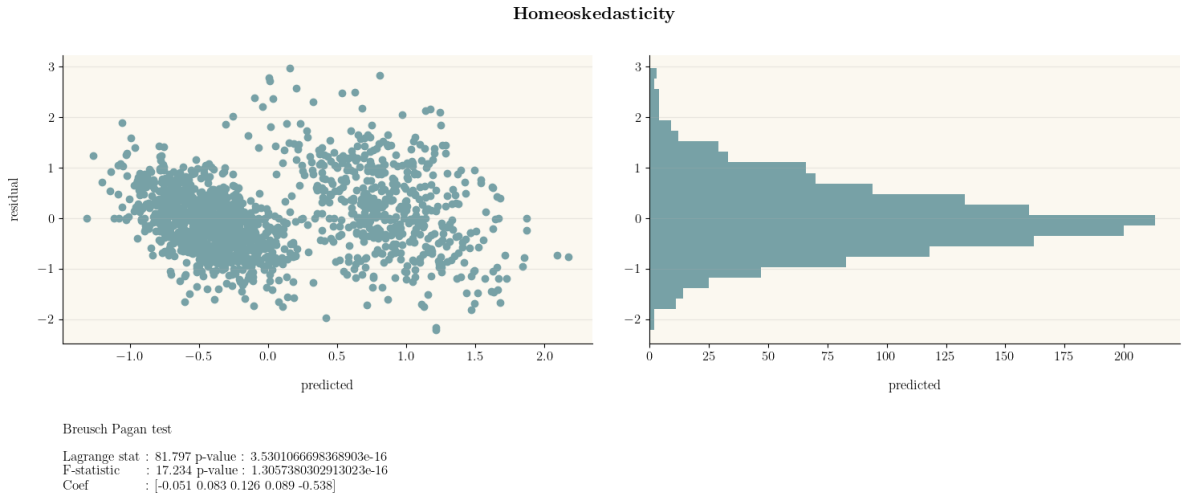



Figure 3: learning curve




Breusch Pagan test

Lagrange stat : 81.797 p-value : 3.5301066698368903e-16
F-statistic : 17.234 p-value : 1.3057380302913023e-16
Coef : [-0.051 0.083 0.126 0.089 -0.538]

Figure 3.1: homeoskedasticity

# 5 Comparaison des modèles

Les modèles ont été comparés à l'aide d'un ensemble de données `test`. En ce qui concerne l'évaluation du modèle non supervisé, une évaluation a été effectuée à la fois avec et sans données de référence (ground_truth).

## 5.1 Performances du modèle de régression logistique

Table 6: Rapport de classification

| **Classe** | **Précision** | **Rappel** | **F1-score** | **Support** |
|---|---|---|---|---|
| Faux billets | 0.99 | 0.99 | 0.99 | 125 |
| Vrais billets | 1.00 | 1.00 | 1.00 | 250 |
| **Précision globale** | | | 0.99 | |
| **Moyenne pondérée** | | | 0.99 | |

Table 7: Matrice de confusion

| | Predict 0 | Predict 1 |
|---|---|---|
| Actual 0 | 124 | 1 |
| Actual 1 | 1 | 249 |

## 5.2 Performance du modèle de clustering k-means

Table 8: Rapport de classification

| **Classe** | **Précision** | **Rappel** | **F1-score** | **Support** |
|---|---|---|---|---|
| Faux billets | 1.00 | 0.98 | 0.99 | 125 |
| Vrais billets | 0.99 | 1.00 | 1.00 | 250 |
| **Précision globale** | | | 0.99 | |
| **Moyenne pondérée** | | | 0.99 | |

Table 9: Matrice de confusion

| | Predict 0 | Predict 1 |
|---|---|---|
| Actual 0 | 123 | 2 |
| Actual 1 | 0 | 250 |

Table 10: Performances sans valeurs de références

| | rand_score |
|---|---|
| truth_compare | 0.989362 |
| expit_compare | 0.989362 |

## 5.3 Visualisation

À l'aide d'une analyse en composantes principales (PCA), une visualisation comparative des modèles a été réalisée, mettant en confrontation leurs performances par rapport aux valeurs de référence. Il ressort que les deux modèles parviennent à discriminer avec une égale efficacité les vrais billets des faux.

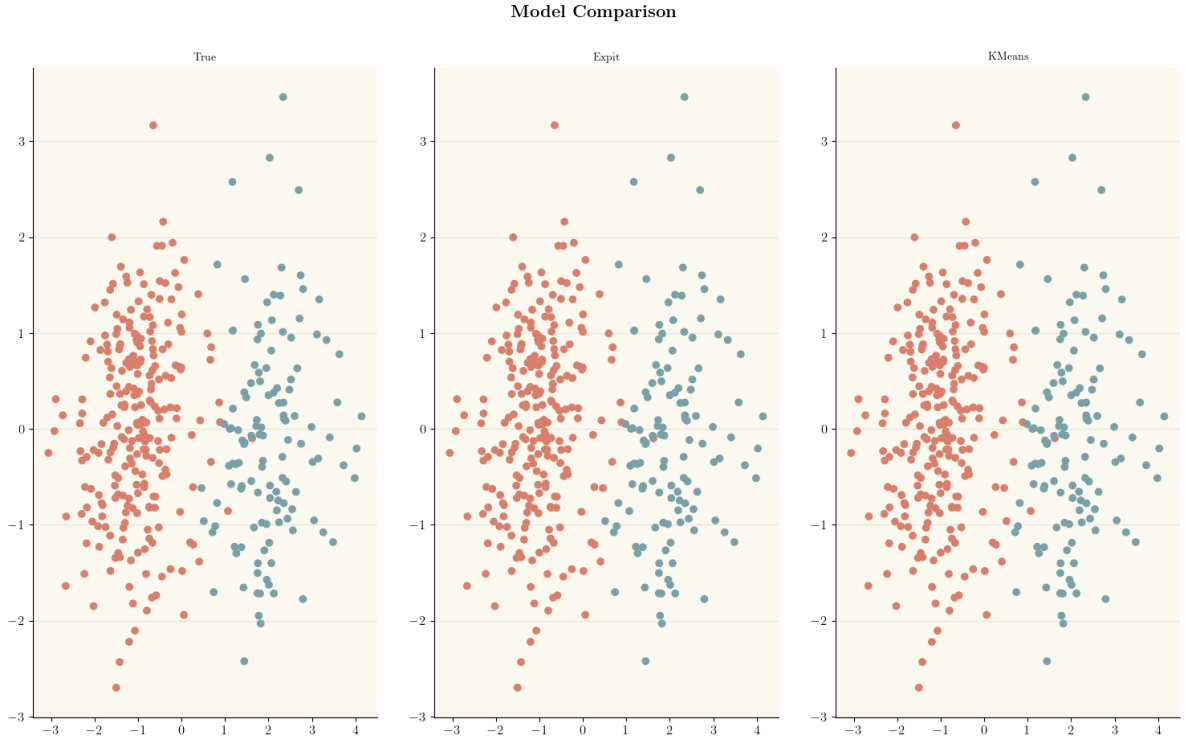



Figure 4: models comparison

## 5.4 Validation

Le jeu de validation est dépourvu de valeurs de référence. Le modèle de régression logistique génère des probabilités d'appartenance aux classes, tandis que pour le modèle de clustering, une analyse des centroïdes est entreprise afin d'identifier ces dernières

Table 11: Régression logistique — probabilités

| **id** | **proba_faux** | **proba_vrai** |
|---|---|---|
| A_1 | 99.96 | 0.04 |
| A_2 | 99.98 | 0.02 |
| A_3 | 99.9 | 0.1 |
| A_4 | 0.62 | 99.38 |
| A_5 | 0.05 | 99.95 |

Table 12: Clustering

| **id** | **cluster** |
|---|---|
| A_1 | 0 |
| A_2 | 0 |
| A_3 | 0 |
| A_4 | 1 |
| A_5 | 1 |

Table 13: Analyse des centroïdes

| | **margin_low** | **margin_up** | **length** |
|---|---|---|---|
| **cluster 0** | 1.12252 | 0.883062 | -1.23536 |
| **cluster 1** | -0.540989 | -0.420671 | 0.5885 |

# 6 Conclusion

Les performances des modèles supervisés et non supervisés sont similaires dans notre étude. Cependant, dans le cadre d'une application de détection de faux billets, un modèle de régression logistique semble plus approprié pour plusieurs raisons :

1. **Interprétabilité :** La régression logistique fournit une interprétation directe des coefficients, ce qui permet de comprendre l'importance de chaque variable dans la prédiction. Cela peut être crucial dans le contexte de la détection de faux billets, où une explication claire des décisions prises par le modèle est nécessaire.

2. **Facilité d'ajustement :** Les modèles de régression logistique sont relativement simples à ajuster et à interpréter. Avec des paramètres bien choisis et un prétraitement approprié des données, il est souvent possible d'obtenir des performances élevées avec ce type de modèle.

En conclusion, bien que les performances des modèles supervisés et non supervisés soient comparables, un modèle de régression logistique offre des avantages significatifs en termes d'interprétabilité et de facilité d'ajustement, ce qui en fait un choix plus approprié pour une application de détection de faux billets.