

Détection automatique de faux billets



Objectif

Préparation des données : les données manquantes

Analyse descriptive

Modélisations : Régression logistique
Kmeans
Random forest

Comparaison et choix du modèle



L'objectif



Identifier des faux billets à partir de six caractéristiques géométriques

- length : la longueur du billet (en mm) ;
- height_left : la hauteur du billet (mesurée sur le côté gauche, en mm) ;
- height_right : la hauteur du billet (mesurée sur le côté droit, en mm) ;
- margin_up : la marge entre le bord supérieur du billet et l'image de celui-ci (en mm) ;
- margin_low : la marge entre le bord inférieur du billet et l'image de celui-ci (en mm) ;
- diagonal : la diagonale du billet (en mm).

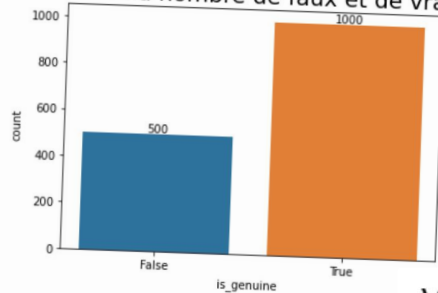
Créer un algorithme capable de les détecter



Les données manquantes

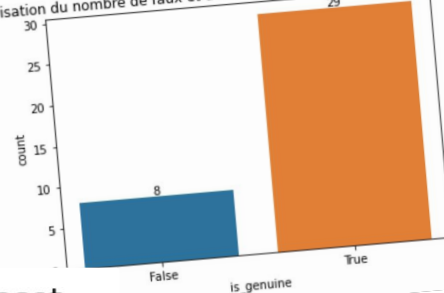


Visualisation du nombre de faux et de vrais billets

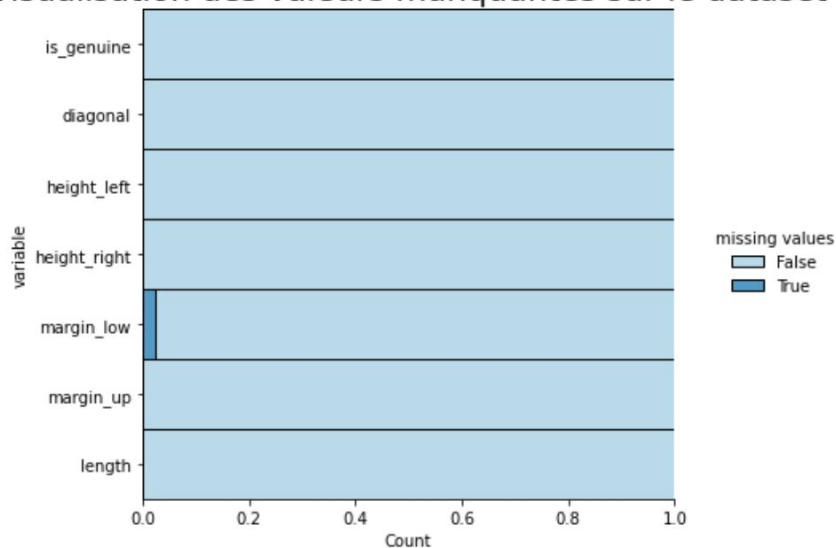


37

Visualisation du nombre de faux et de vrais billets pour les valeurs manquantes

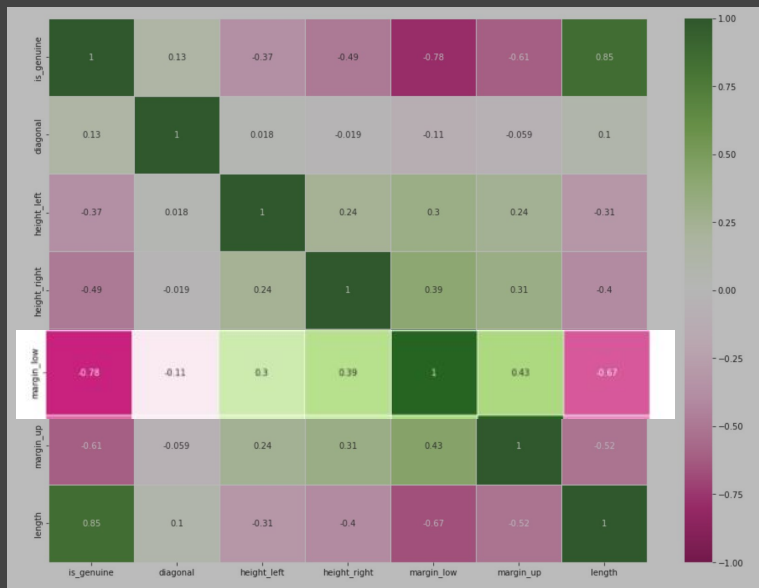


Visualisation des valeurs manquantes sur le dataset



Pourcentage de données manquantes 2.47 %

Imputation des données manquantes par régression linéaire (variables quantitatives)



(all $p < 0.05$)

Linéarité des relations entre variables explicatives et margin_low

Régression linéaire simple (une seule variable explicative de margin_low)

Régression linéaire multiple (plusieurs variables explicatives)

MODÈLE COMPLET



MODÈLE SIMPLIFIÉ

MODÈLE COMPLET

| | | | |
|-------------------|------------------|---------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.615 |
| Method: | Least Squares | F-statistic: | 390.7 |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 4.75e-299 |
| Time: | 11:12:40 | Log-Likelihood: | -774.14 |
| No. Observations: | 1463 | AIC: | 1562. |
| Df Residuals: | 1456 | BIC: | 1599. |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|--------------|---------|---------|---------|-------|---------|--------|
| Intercept | 2.8668 | 8.316 | 0.345 | 0.730 | -13.445 | 19.179 |
| is_genuine | -1.1406 | 0.050 | -23.028 | 0.000 | -1.238 | -1.043 |
| diagonal | -0.0130 | 0.036 | -0.364 | 0.716 | -0.083 | 0.057 |
| height_left | 0.0283 | 0.039 | 0.727 | 0.468 | -0.048 | 0.105 |
| height_right | 0.0267 | 0.038 | 0.701 | 0.484 | -0.048 | 0.102 |
| margin_up | -0.2128 | 0.059 | -3.621 | 0.000 | -0.328 | -0.098 |
| length | -0.0039 | 0.023 | -0.166 | 0.868 | -0.050 | 0.042 |

| | | | |
|----------------|--------|-------------------|----------|
| Omnibus: | 21.975 | Durbin-Watson: | 2.038 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 37.993 |
| Skew: | 0.061 | Prob(JB): | 5.62e-09 |
| Kurtosis: | 3.780 | Cond. No. | 1.95e+05 |

MODÈLE RÉDUIT

| | | | |
|-------------------|------------------|---------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 1174. |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 1.24e-304 |
| Time: | 11:12:40 | Log-Likelihood: | -774.73 |
| No. Observations: | 1463 | AIC: | 1555 |
| Df Residuals: | 1460 | BIC: | 1571. |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|------------|---------|---------|---------|-------|--------|--------|
| Intercept | 5.9263 | 0.198 | 30.003 | 0.000 | 5.539 | 6.314 |
| is_genuine | -1.1632 | 0.029 | -40.477 | 0.000 | -1.220 | -1.107 |
| margin_up | -0.2119 | 0.059 | -3.612 | 0.000 | -0.327 | -0.097 |

| | | | |
|----------------|--------|-------------------|----------|
| Omnibus: | 22.365 | Durbin-Watson: | 2.041 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 39.106 |
| Skew: | 0.057 | Prob(JB): | 3.22e-09 |
| Kurtosis: | 3.793 | Cond. No. | 65.0 |

Après sélection descendante (backward selection), ce modèle de régression linéaire multiple est retenu

Examen des conditions de validité du modèle

MODÈLE RÉDUIT

| | | | |
|----------------|---------------|-----------------|-------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.615 |
| Method: | Least Squares | F-statistic: | 390.7 |

| | | | |
|-------------------|------------------|---------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 1174. |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 1.24e-304 |
| Time: | 11:12:40 | Log-Likelihood: | -774.73 |
| No. Observations: | 1463 | AIC: | 1555 |
| Df Residuals: | 1460 | BIC: | 1571. |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|------------|---------|---------|---------|-------|--------|--------|
| Intercept | 5.9263 | 0.198 | 30.003 | 0.000 | 5.539 | 6.314 |
| is_genuine | -1.1632 | 0.029 | -40.477 | 0.000 | -1.220 | -1.107 |
| margin_up | -0.2119 | 0.059 | -3.612 | 0.000 | -0.327 | -0.097 |

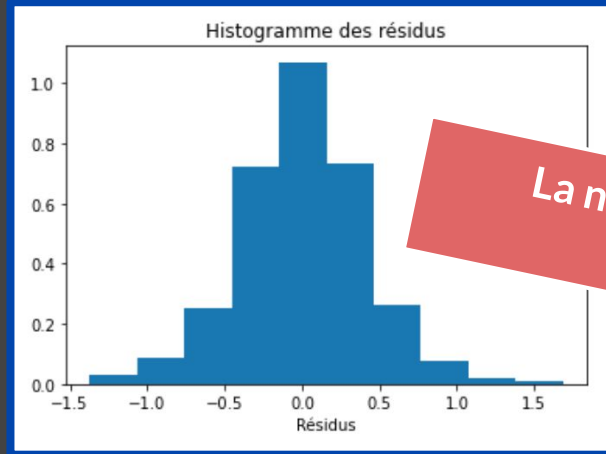
| | | | |
|----------------|--------|-------------------|----------|
| Omnibus: | 22.365 | Durbin-Watson: | 2.041 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 39.106 |
| Skew: | 0.057 | Prob(JB): | 3.22e-09 |
| Kurtosis: | 3.793 | Cond. No. | 65.0 |

| | | | |
|----------------|-------|-------------------|----------|
| | | Durbin-Watson: | 2.038 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 37.993 |
| Skew: | 0.061 | Prob(JB): | 5.62e-09 |
| Kurtosis: | 3.780 | Cond. No. | 1.95e+05 |

| | | | |
|--------------------------|------------------|----------------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 1174. |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 1.24e-304 |
| Time: | 11:12:40 | Log-Likelihood: | -774.73 |
| No. Observations: | 1463 | AIC: | 1555. |
| Df Residuals: | 1460 | BIC: | 1571. |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|-------------------|---------|---------|---------|-------|--------|--------|
| Intercept | 5.9263 | 0.198 | 30.003 | 0.000 | 5.539 | 6.314 |
| is_genuine | -1.1632 | 0.029 | -40.477 | 0.000 | -1.220 | -1.107 |
| margin_up | -0.2119 | 0.059 | -3.612 | 0.000 | -0.327 | -0.097 |

| | | | |
|-----------------------|--------|--------------------------|----------|
| Omnibus: | 22.365 | Durbin-Watson: | 2.041 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 39.106 |
| Skew: | 0.057 | Prob(JB): | 3.22e-09 |
| Kurtosis: | 3.793 | Cond. No. | 65.0 |

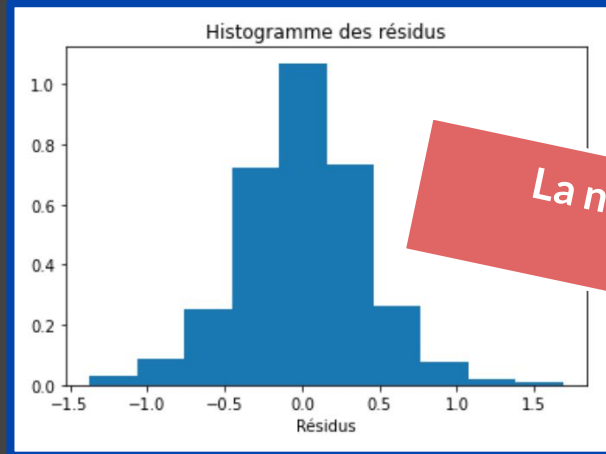


La normalité des résidus est rejetée

| | | | |
|--------------------------|------------------|----------------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 1174. |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 1.24e-304 |
| Time: | 11:12:40 | Log-Likelihood: | -774.73 |
| No. Observations: | 1463 | AIC: | 1555. |
| Df Residuals: | 1460 | BIC: | 1571. |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|-------------------|---------|---------|---------|-------|--------|--------|
| Intercept | 5.9263 | 0.198 | 30.003 | 0.000 | 5.539 | 6.314 |
| is_genuine | -1.1632 | 0.029 | -40.477 | 0.000 | -1.220 | -1.107 |
| margin_up | -0.2119 | 0.059 | -3.612 | 0.000 | -0.327 | -0.097 |

| | | | |
|-----------------------|--------|--------------------------|----------|
| Omnibus: | 22.365 | Durbin-Watson: | 2.041 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 39.106 |
| Skew: | 0.057 | Prob(JB): | 3.22e-09 |
| Kurtosis: | 3.793 | Cond. No. | 65.0 |



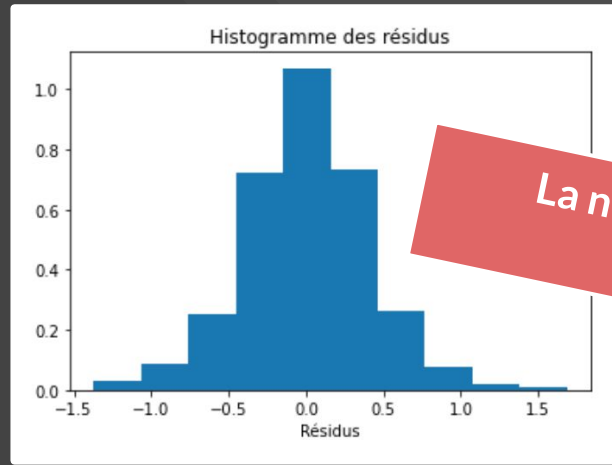
La normalité des résidus est rejetée

L'homoscédasticité (égalité des variances) est rejetée
Test de Levene ($p < 0,05$)

| | | | |
|--------------------------|------------------|----------------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 1174. |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 1.24e-304 |
| Time: | 11:12:40 | Log-Likelihood: | -774.73 |
| No. Observations: | 1463 | AIC: | 1555. |
| Df Residuals: | 1460 | BIC: | 1571. |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|-------------------|---------|---------|---------|-------|--------|--------|
| Intercept | 5.9263 | 0.198 | 30.003 | 0.000 | 5.539 | 6.314 |
| is_genuine | -1.1632 | 0.029 | -40.477 | 0.000 | -1.220 | -1.107 |
| margin_up | -0.2119 | 0.059 | -3.612 | 0.000 | -0.327 | -0.097 |

| | | | |
|-----------------------|--------|--------------------------|----------|
| Omnibus: | 22.365 | Durbin-Watson: | 2.041 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 39.106 |
| Skew: | 0.057 | Prob(JB): | 3.22e-09 |
| Kurtosis: | 3.793 | Cond. No. | 65.0 |



La normalité des résidus est rejetée

L'homoscédasticité (égalité des variances) est rejetée
Test de Levene ($p < 0,05$)

Pas de multicollinéarité
($VIF < 10$)

Après sélection descendante
(backward selection), et
vérification des conditions de
validité du modèle réduit, les
données manquantes sur la
variable 'margin_low' sont
imputées à partir de ce modèle
de régression linéaire multiple

fonction utilisée : predict()

MODÈLE RÉDUIT

| | | | |
|-------------------|------------------|---------------------|-----------|
| Dep. Variable: | margin_low | R-squared: | 0.617 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 1174. |
| Date: | Wed, 06 Apr 2022 | Prob (F-statistic): | 1.24e-304 |
| Time: | 11:12:40 | Log-Likelihood: | -774.73 |
| No. Observations: | 1463 | AIC: | 1555 |
| Df Residuals: | 1460 | BIC: | 1571. |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|------------|---------|---------|---------|-------|--------|--------|
| Intercept | 5.9263 | 0.198 | 30.003 | 0.000 | 5.539 | 6.314 |
| is_genuine | -1.1632 | 0.029 | -40.477 | 0.000 | -1.220 | -1.107 |
| margin_up | -0.2119 | 0.059 | -3.612 | 0.000 | -0.327 | -0.097 |

| | | | |
|----------------|--------|-------------------|----------|
| Omnibus: | 22.365 | Durbin-Watson: | 2.041 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 39.106 |
| Skew: | 0.057 | Prob(JB): | 3.22e-09 |
| Kurtosis: | 3.793 | Cond. No. | 65.0 |



Description des données



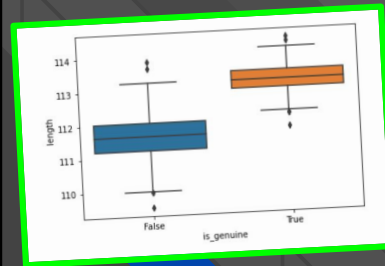
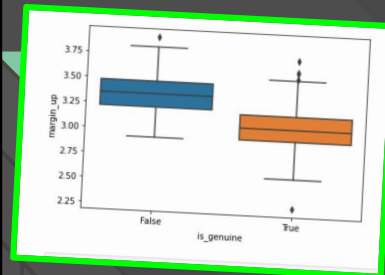
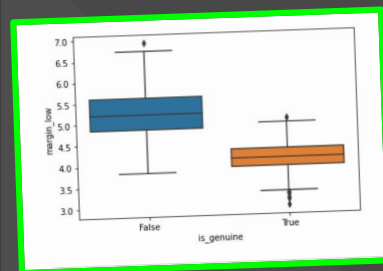
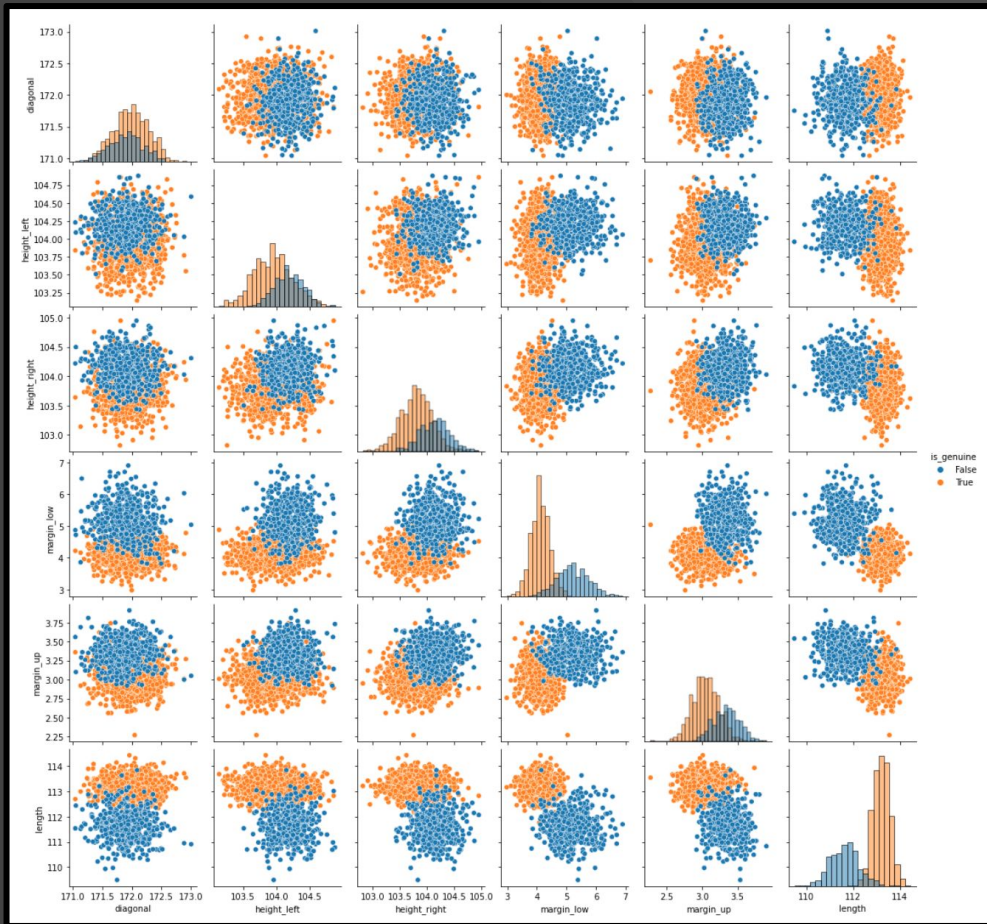
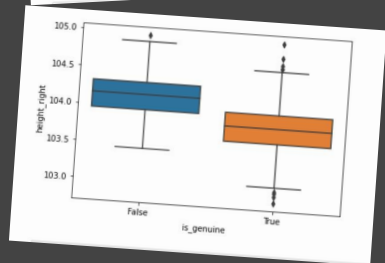
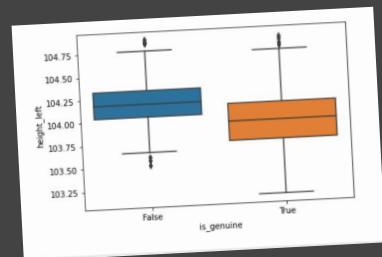
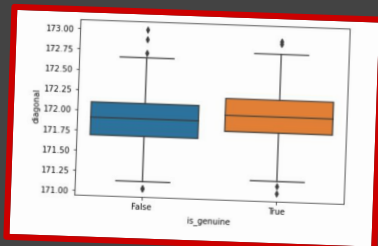
1000



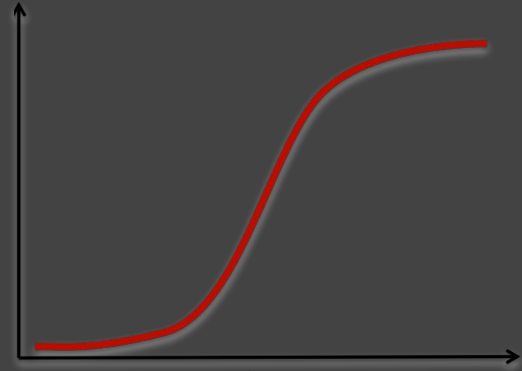
500



- longueur
 - hauteur gauche
 - hauteur droite
 - marge supérieure
 - marge inférieure
 - diagonal
- (en mm)

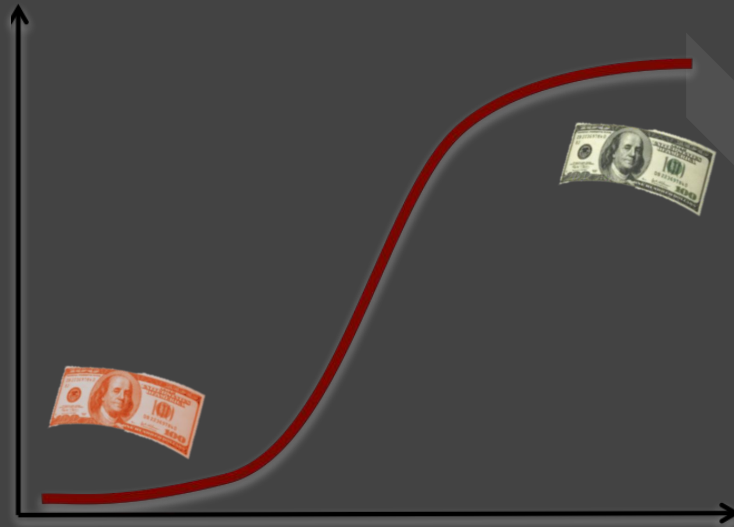


Modélisation : La régression logistique

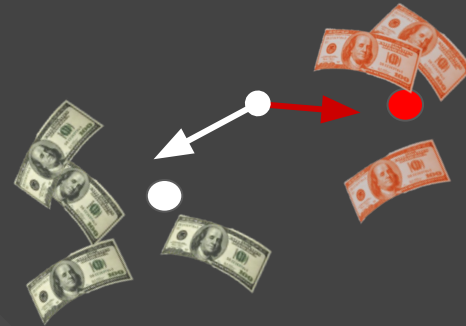


Comme une régression linéaire, une **régression logistique** permet d'étudier la relation entre des **variables explicatives** et une **variable booléenne** mais en utilisant **une fonction logistique**.

→ Prédire la probabilité qu'un billet soit vrai ou faux selon les dimensions



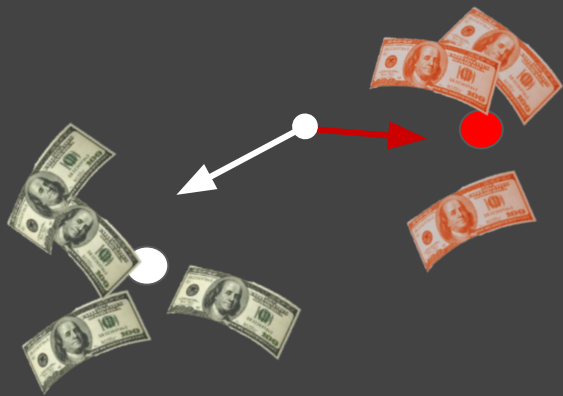
Modélisation : La méthode Kmeans



L'algorithme Kmeans

Déterminer le nombre de clusters voulus (méthode du coude) 2

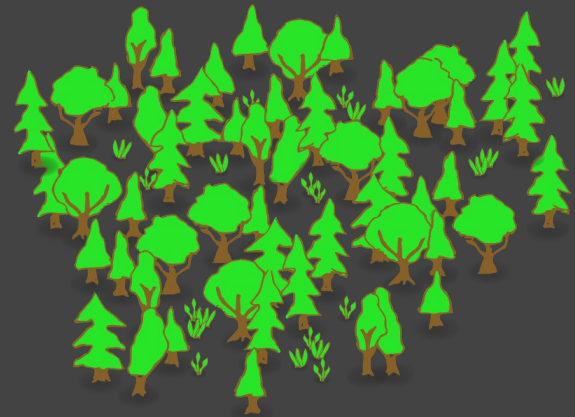
L' algorithme choisit arbitrairement 2 centroïdes et mesure la distance de chaque individu aux différents centroïdes pour leur attribuer un centroïde






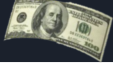
Dispersion (l'inertie) intra groupe

Inertie inter groupe

Modélisation : La méthode Random Forest





Pour classer : construction de plusieurs arbres à partir des données en prenant des individus du dataset d'entraînement (échantillon aléatoire avec remise) et quelques variables pour chaque arbre.


| | var 1 | var 2 | var ... | Y |
|------|-------|-------|---------|---|
| ind1 | | | |  |
| ind2 | | | |  |
| ind3 | | | |  |
| ind4 | | | |  |

| | | | | |
|------|--|--|--|----------|
| test | | | | ? |
|------|--|--|--|----------|

Bootstrapped dataset

|  | var 3 | var 4 |
|--|-------|-------|
| 2 | | |
| 2 | | |
| 3 | | |
| 3 | | |

|  | var 1 | var 3 |
|---|-------|-------|
| 1 | | |
| 3 | | |
| 4 | | |
| 4 | | |

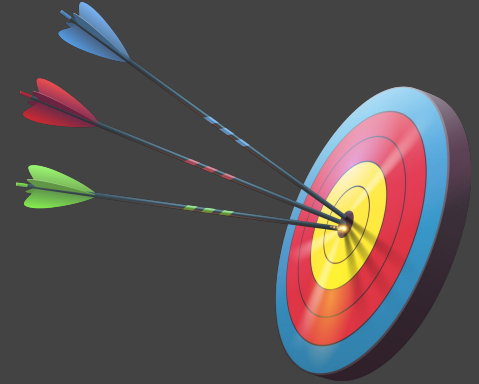
|  | var 2 | var 3 |
|---|-------|-------|
| 1 | | |
| 2 | | |
| 3 | | |
| 3 | | |

BAGGING

Agrégation



Modélisation : Comparaison et choix du modèle



ÉVITER QUE LES FAUX BILLETS SOIENT
IDENTIFIER EN VRAIS BILLETS

Les mesures de **performance** d'un modèle

F1 score

Combien d'individus ont été correctement prédits ?

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$



Précision

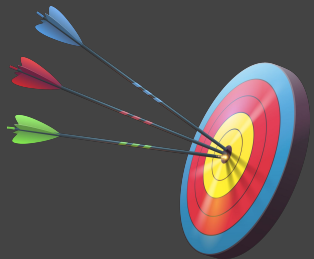


Rappel

PRÉCISION
 $TP / (TP + FP)$

RECALL
 $TP / (TP + FN)$

| | | PREDICTION | |
|-------------|---|------------------------------------|-----------------------------------|
| | | FAUX | VRAI |
| OBSERVATION | FAUX  | true negative TN | false positive FP TYPE I ERROR |
| | VRAI  | false negative FN TYPE II ERROR | true positive TP |



Les matrices de confusion

Régression logistique

Performance : 99,73

Kmeans

Performance : 97,86

Random forest

Performance : 99,73

| Predictions | False | True |
|--------------|-------|------|
| Observations | | |
| False | 125 | 1 |
| True | 0 | 249 |

| Predictions | 0 | 1 |
|--------------|-----|-----|
| Observations | | |
| False | 122 | 4 |
| True | 4 | 245 |

| Predictions | False | True |
|--------------|-------|------|
| Observations | | |
| False | 125 | 1 |
| True | 0 | 249 |

Testons l'algorithme





Merci

