1) Sommaire

* Introduction
* Analyses univariées et bivariées
* PCA
* Apprentissage non-supervisé
* Apprentissage supervisé
* Bilan

2) Introduction  
  
*Objectif : Créer un algorithme de vérification des billets performant avec un Taux de réussite > 99%*

*3)Je commence par faire une analyse univarié*

*J’analyse les courbe de densité de mes 6 variables :*

* la longueur du billet (en mm) ; length
* la hauteur du billet (mesurée sur le côté gauche, en mm) ; height\_left
* La hauteur du billet (mesurée sur le côté droit, en mm) ; height\_right
* la marge entre le bord supérieur du billet et l'image de celui-ci (en mm) ; margin\_up
* la marge entre le bord inférieur du billet et l'image de celui-ci (en mm) ; margin\_low
* la diagonale du billet (en mm). diagonal

*Mes variable semble distribué normalement sauf length et margin\_low.*

*Length semble former deux gaussienne,  potentiellement 2 sous groupe distribués normalements (cette variable à de forte chance d’être intéressante pour la suite)*

4)Je décide ensuite de faire un shapiro test pour confirmer mes analyses

Le Shapiro test confirme, les p-value sont supérieur à 0.05 sauf pour margin\_low et length.

mes variables semble donc distribué normalement sauf la longueur et la marge du bas

5)Je passe à l’analyse bivarié

Je construit la matrice de corrélation de mes variables pour comprendre les relations entre elles

J’observe que mes variable height\_right et height\_left sont corrélées r² = 0.53

height\_right et margin\_low sont légerement corrélé r² = 0.26

Length et margin\_low sont anti corrélé r²=0.41

length et margin up sont légèrement anti-corrélé r² = 0.28

6)Je trace les courbe de mes paires de variable corrélé et anti corrélé pour mieu comprendre les relations.

Dans mon premier graphique je vois simplement deux variables très corrélé. Height\_right et height\_left

Dans le deuxième cela semble moins évident.

Les deux dernier graphique sont particulièrement intéressant car 2 groupes semble apparaitre.

(sans doute les vrai et les faux billet, les variable length et margin vont surement être de très bonne variable explicative)

7) Je décide de faire une pca pour pouvoir mieux analyser mes données.

Je commence par analyser mon ébloui des valeurs propres afin de choisir le nombre de plan à étudier.

Je décide de n’étudier que mes deux premiers axes pour plusieurs raisons :

* ils sont supérieur à la moyenne de variance expliqué par axe 16.7% (elles respecte donc le critère de kaiser)
* Il y a un décrochage à partir du troisième axe (Le critère du coude nous suggère aussi de n’utiliser que les 2 premiers axes)

Je décide donc d’analyser mes valeurs seulement dans mon premier plan factoriel formé des axes 1 et 2 même si il ne représente que 69.4% de l’inertie totale

8)J’analyse maintenant le cercle de corrélation de ma pca

Pour simplifier drastiquement je peux dire que l’axe des y va dépendre de la diagonal plus la diagonal est important plus le billet sera haut alors que l’axe des x va dépendre des 5 autres variables les hauteur et les marges sont corrélé positivement à l’axe des x alors que la longueur est corrélé négativement. (Je retrouve les corrélations évoqué précédemment sur ce graphique)

9)J’analyse le premier plan factoriel de ma pca

Mes billet sont séparé en deux groupes un à gauche un à droite qui correspondent assez bien à mes groupes vrai et faux.

J’ai l’impression que la variable diagonal n’a pas d’intérêt pour déterminer la nature de mes billets elle semble juste éparpiller mes deux groupes.

Au vue de la disposition des centroide(le vrais en haut à gauche le faux en bas à droite j’ai l’impression que les trois variables qui permettent de différencier le mieux mes billet sont la longueur et les marges (ceci n’est qu’une supposition)

10) je vais étudier la contribution de mes individus aux axe pour vérifier que ma pca est pertinente

1er axe :

La participation maximal est de 2.7% à peu près 4 fois la moyenne. Aucun individu n’a une participation assez forte pour fausser mon acp .

Les 4 individus qui participent le plus fortement a mon axe principale sont les billets 123, 50, 30 et 113 (2 faux 2 vraies)

11) 2em axe :

la participation maximal est de 4% à peu près 8 fois la moyenne. Là encore Aucun individu n’a une participation assez forte pour fausser mon acp .

Mon acp semble donc être de qualité de ce point de vue là.

Les 2 individus qui participent le plus fortement a mon axe principale sont les billets 6 et 167 (1 faux 1 vraies)

12) J’effectue un kmeans clustering à deux clusters sur l’ensemble de mes données.

Ici mes deux cluster se chevauche sur mon premier plan factoriel, en général on préfère qu’il soit bien séparés. Mais vu que le premier plan factoriel représente moins de 70% de l’inertie totale, il se peut que notre classification soit quand même très bonne.

Afin de mieux pouvoir juger notre classification je vais tout simplement comparer mes 2 clusters obtenues au deux cluster préexistant faux billets et vraie billet

(si ils correspondent c’est que la classification est très intéressante et que mes billets sont facilement contrôlable. Au contraire si ils ne correspondent pas je devrai trouver un meilleur moyen de classer mes billets (possiblement avec de nouvelle variable..) ou avec un autre algorithme, peut être avec un algorithme supervisé.)

13)J’obtient que ma classification correspond à 98,2 % à la classification faux billet vrai billets. Si j’avais utilisé mon algorithme pour vérifier que mes billets sont vrais ou faux je ne me serai trompé qu’a trois reprises sur 170 billets.

Même si cela est déjà très bien je vais tout de même essayer de classer mes billets avec d’autres méthodes pour voir si cela peut augmenter mon pourcentage.

14) Je décide d’utiliser un algorithme d’apprentissage supervisé, je fais une regression logistique sur les 6 variables

J’évalue le taux de réussite de mon algorithme en réalisant 1000 cross validation 70/30 sur mes données.

J’obtiens un résultat inférieur à celui trouvé à l’aide d’un kmeans clustering, je décide donc de supprimer des variables pour augmenter mon taux de réussite.

J’ai déjà a ce stade une bonne idée des variables à supprimer (diagonal et 1 hauteur).

En effet on a vu dans la pca que la variable diagonal ne faisait que disperser mes groupes et n’était pas une bonne variable prédictive .

Et on a vu que mes deux hauteurs sont fortement corrélé donc en supprimer une devrait être bénéfique à l’algorithme

Mais je vais utiliser un test pour confirmer mes idée.

15) Je vais utiliser le critère d’information d’Akaike soit l’AIC,

(Le critère d'information d'Akaike permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le [critère de parcimonie](https://fr.wikipedia.org/wiki/Rasoir_d%27Ockham). On choisit alors le modèle avec le critère d'information d'Akaike le plus faible)

L’AIC le plus faible est obtenu en supprimant les variables hauteur et diagonal.

Cela correspond à peu près à ce a quoi je m’attendais

je décide donc de retirer ces variable et de garder seulement la longueur et les marges

16) Je réalise donc une deuxieme regression logistique cette fois seulement sur les trois variables présélectionnées.

Mon taux de réussite a augmenté je suis maintenant au dessus du taux obtenue avec un kmeans clustering. Mais la différence est très faible et je voudrais au moins 99% de réussite

Je décide alors de changer de méthode

17) J’utilise cette fois une LDA analyse discriminante linéaire.

La **LDA** est une **méthode** d'analyse numérique qui permet de chercher la combinaison linéaire des variables qui sépare au mieux les données.

((Elle permet de maximiser l'éparpillement inter-classes et de réduire l'éparpillement intraclasses.))

J’obtient un taux de réussite de 99,3 % !

(on peut voir que les coefficient les plus important de ma LDA sont sur les variables que le test AIC avait conservé, les marges et la longueur)

18) Au vue de mon échantillon de billet, la meilleur méthode parmi celles que

j’ai testé est l’Analyse discriminante linéaire. Elle atteint bien l’objectif que j’avais ciblé,

avoir un taux de réussite supérieur à 99%

J’ai créé une application web qui permet de tester des billets à l’aide de cette méthode mais aussi de la régression logistique.

19) Ce projet m’a permis de me familiariser avec les différents types d’apprentissages supervisés et non supervisés et il m’a aussi donné l’occasion de créer un programme.

Le projet demandait d’utiliser seulement la régression logistique et je trouve ça dommage..