# Détectez des faux billets avec Python

## Analyses descriptive des données

**12 .** Heatmap des corrélations :

**Les 2 variables quantitative les plus corréler sont 'length' et 'is_genuine',**

**elle sont corréler Positivement !**

'length' est aussi assez bien corréler négativement à 'margin_low' et 'margin_up'

'is_genuine' est aussi assez bien corréler à 'margin_low' et 'margin_up'

Liste des variables les plus corréler à 'is_genuine' dans l'ordre décroissant :

- length (0.85, positif)
- margin_low (-0.78, négatif)
- margin_up (-0.61, négatif)
- height_right (-0.49, négatif)
- height_left (-0.38, négatif)
- diagonal (0.13, positif)

Variable à forte corrélation avec les autres :

- length

Variable à faible corrélation :

- diagonal

## **13 .** Relation entre les variables :

( Visualisation par nuage de points )

**Les 2 groupes sont le plus distinct sur la ligne de la variable 'length'**

Groupe 1 :

is_genuine et length = Forte corrélation positive (0.85). Ces deux variables évoluent ensemble : quand length augmente, is_genuine tend à augmenter également.

Groupe 2 :

is_genuine et margin_low = Forte corrélation négative (-0.78) Ces variables sont opposées : quand margin_low augmente, is_genuine diminue fortement.

## **14 .** Comparaison visuelle des vrais et faux billets pour chaque variables:

( Boîte à moustaches )

**Les variables 'length' et 'margin_low' et 'margin_up' ont le plus de différence entre les vrais et les faux billets !**

Length = Corrélation positive (0.85) | Les billets authentiques ont tendance à avoir une longueur différente des faux billets.

Margin_low = Corrélation négative (-0.78) | Une faible marge basse est souvent associée aux vrais billets, alors qu'elle est plus grande sur les faux.

Margin_up = Corrélation négative (-0.61) | Une faible marge haute est plus fréquente sur les vrais billets que sur les faux.

**15 .** Visualisation des vrais/faux billets sur margin_low par rapport à length :

Certains paramètres montrent une grande variabilité entre les groupes,

ce qui pourrait indiquer des caractéristiques clés permettant de distinguer les vrais des faux billets.

On observe des différences marquées sur des variables comme la longueur et les marges,

qui ont déjà été identifiées comme des facteurs importants de distinction.

---

**Enrichissement des données**

**16 .** Visualisation de la droite de Régression linéaire :

La régression linéaire multiple prend en compte plusieurs variables donc plusieurs dimensions.

Ici nous avons 6 variables donc 6 dimensions !

Malheureusement on ne peut pas visualiser plus de 3 dimensions !

**17.** Visualisation des valeurs prédis de la droite de Régression linéaire :

Les valeurs de 'margin_low' prédites grace à la régression linéaire multiple avec nos données, comme on peut le voir sur cet exemple !

---

**Analyses en Composantes Principales**

**18 .** Normalisation des données et Variance expliquée :

Les données sont standardisées pour faciliter leurs études !

**19 .** Cercle des Corrélations et Projections des Individus :

Lecture du cercle des corrélations

Les 2 variables les mieux représenter sont 'diagonal' et 'length', car les flèches de ces variables sont les plus grandes !

- L'axe des ordonnées est représenter par 'diagonal'

- L'axe des abscisses est représenter par 'length'

Lecture de la Projection des Individus

On constate que l'on distingue bien nos 2 groupes (vrai billets et faux billets) sur nos 2 premières composantes !

- Les vrais billets se concentrent sur le côté gauche

- Les faux billets se concentrent sur le côté droit

- Nous voyons une zone "d'incertitude" qui est la zone de contact entre les 2 groupes

Les billets sont le plus différencier grâce à leur longueur (length), car c'est cette variable qui est le mieux représenter sur l'axe des abscisses !

---

**Classification supervisée**

**20 .** K-means ( Méthode du Coude ) :

On observe sur le graphique, que la plus grosse cassure se trouve au niveau de 2 clusters !

**21 .** K-means ( Score de Silhouette ) :

On observe sur le graphique, que l'on obtient le plus gros score à 2 clusters !

On choisit donc d'utiliser le nombre de 2 clusters pour le K-Means !

**22 .** K-Means (Visualisation du modèle) :

On constate bien nos 2 groupes sur nos 2 premières composantes !

- Les centroïdes crée par le K-Means sont proche à ceux des données de base

- Les groupes sont très ressemblant à ceux des données de base

- La zone "d'incertitude" n'existe plus ! On voit une limite entre les 2 groupes !

**23 .** K-Means (Evaluation du modèle) :

La matrice de confusion nous permet d'évaluer les résultats du K-Means par rapport à nos données de base !

- 156 billets ont été considérés comme vrais alors qu'ils étaient faux (31.52% des données)

- 333 billets ont été considérés comme faux alors qu'ils étaient vrais (67.27% des données)

- 98.79% de prédictions incorrectes ! = somme des erreurs 31.52% + 67.27% = 98.79% (taux total d'erreur, qui montre que le modèle K-Means n'a presque aucune précision dans cette tâche. Seuls 1.21% des billets ont été correctement classés !)

*Accuracy* : Cet indicateur nous indique le pourcentage de prédictions correctes que notre modèle a réalisé sur l'ensemble de données de test. Plus la précision est élevée, meilleur est notre modèle.

inAccuracy : 98.79% de prédictions incorrectes !

et une Accuracy de 1.21% de prédictions correctes !