Qu’est-ce qu’une ACP (Analyse en Composantes Principales) ?

C’est une technique de réduction de dimension : elle transforme des variables initiales (souvent corrélées) en nouvelles variables indépendantes appelées composantes principales,

tout en conservant un maximum d’information.

Pourquoi l’utiliser dans mon projet ?

- Les variables comme “`diagonal`”, “`height_left`”, “`height_right`”, etc. sont très corrélées.
- L’ACP permet de projeter les billets sur un plan pour visualiser leur répartition.
- Cela aide à détecter des groupes ou anomalies, même avant d’entraîner un modèle.

J’ai utilisé l’ACP pour mieux visualiser la structure des billets dans l’espace.

Cela m’a permis d’identifier des regroupements naturels entre billets vrais et faux, avant même d’entraîner un modèle.

Elle aide aussi à vérifier si les variables sont redondantes ou si les données sont bien séparables.

Classification supervisée?

Une classification supervisée apprend à partir d’exemples déjà classés (ex : billet vrai ou faux).

Elle est utilisée pour prédire une étiquette / classe.

Classification non supervisée ?

une classification non supervisée cherche à découvrir des structures cachées sans connaître la réponse à l’avance.

J’ai utilisé les deux approches dans le projet :

- l’**ACP et le KMeans** pour explorer et visualiser
- la **régression logistique** pour construire un modèle prédictif.

Un cluster, dans un modèle non supervisé :

- C’est un groupe découvert automatiquement par l’algorithme.
- Les clusters sont créés sans connaître les classes réelles.
- Ce n’est pas une étiquette, c’est une structure trouvée automatiquement.

KMeans est-il supervisé ou non supervisé ? Pourquoi ?

KMeans est un algorithme de classification non supervisée.

Il regroupe les données en clusters en fonction de leurs similarités (distances).

Il n'utilise aucune étiquette connu**e.**

J'ai utilisé KMeans pour voir si les billets se regroupaient naturellement entre vrais et faux, sans donner les étiquettes.

Cela m'a permis de valider visuellement que certaines variables étaient discriminantes.

Ce n'est pas un algorithme prédictif ici, mais exploratoire.

La régression linéaire est-elle supervisée ou non supervisée ?

La régression linéaire est un modèle supervisé.
Elle prédit une valeur continue à partir de variables explicatives.
Exemple : prédire un prix, une taille, une température..

Ce n'est pas adaptée à un problème comme le mien, où il faut prédire une classe binaire (vrai ou faux).

C'est pourquoi j'ai utilisé la régression logistique, qui est aussi supervisée, mais adaptée à la classification.

Comment ai-je séparé mes données avec la régression logistique ?

J'ai séparé mes données en jeu d'entraînement et jeu de test.

- Entraînement du modèle sur le jeu d'entraînement avec `fit()`.
- Prédictions réalisées sur le jeu de test.
- Concaténation des résultats avec les données originales pour interpréter les prédictions (id, proba, vrai/faux).

J'ai d'abord séparé mes données avec `train_test_split()` pour entraîner mon modèle de manière fiable. Cela garantit que mon modèle généralise bien, et qu'il ne mémorise pas seulement les données d'entraînement.
Une bonne séparation des données est indispensable dans un apprentissage supervisé fiable.

Ensuite, j'ai utilisé la régression logistique pour prédire la probabilité qu'un billet soit vrai.
J'ai ensuite concaténé les prédictions et les probabilités avec les données initiales pour avoir une vue claire et exploitable :

on y retrouve l'identifiant du billet, la probabilité d'authenticité, et le résultat final (vrai ou faux billet).

Cela permet de :
- Entraîner le modèle sans biais
- Évaluer sa performance sur des données jamais vues

Corrélation

La **corrélation** mesure la **force** et la **direction** de la relation linéaire entre deux variables numériques.

- **Valeurs possibles :** entre **-1** et **+1**
    - **+1** → corrélation parfaitement **positive** (quand l'une augmente, l'autre aussi)
    - **0** → **aucune** corrélation linéaire
    - **-1** → corrélation parfaitement **négative** (quand l'une augmente, l'autre diminue)

  📊 Exemple : Il y a une forte corrélation positive entre la taille d'une personne et son poids.

Types de corrélations :

- **Positive** : les deux variables évoluent dans le même sens.
- **Négative** : une variable augmente quand l'autre diminue.
- **Nulle** : pas de lien linéaire évident.

Centroïde

En **apprentissage automatique** (machine learning), un **centroïde** est le **centre géométrique** d'un groupe de données.

- Principalement utilisé dans des **algorithmes de regroupement (clustering)**, comme **K-Means**.
- Le centroïde représente le **point moyen** (en coordonnées) de toutes les données appartenant à un **cluster**.

Exemple : Si un cluster regroupe des clients aux revenus similaires, le centroïde représente le "client moyen" en termes de revenu et autres caractéristiques.

En pratique :

Après avoir divisé les données en **K clusters**, on calcule le **centroïde de chaque cluster**.

Ensuite, chaque point est **rattaché au centroïde le plus proche**, et on réajuste les centroïdes jusqu'à convergence.

Analyses en Composantes Principales

18 . Normalisation des données et Variance expliquée :

Les données sont standardisées pour faciliter leurs études !

Avant d'entraîner mes modèles de classification pour détecter les vrais et faux billets, j'ai normalisé les données.

C’est une étape cruciale, car les différentes caractéristiques des billets — comme la hauteur, la largeur ou la marge — n’ont pas les mêmes unités ni les mêmes échelles.

Par exemple, la hauteur d’un billet peut aller jusqu’à 100 mm, alors qu’une marge latérale peut mesurer seulement quelques millimètres.

Si je ne normalise pas, la hauteur va dominer les calculs, simplement parce que ses valeurs sont plus grandes. Cela peut fausser les résultats, notamment pour les algorithmes sensibles aux distances.

Grâce à cette ACP, j’ai obtenu un **éboulis des variances expliquées**, qui m’indique **à elle seule 60,2% de l'information totale du jeu de données**.

Cela signifie que cette seule composante résume déjà une grande partie des variations entre les billets, ce qui est très utile pour simplifier le problème tout en gardant une bonne représentation.

La normalisation était donc indispensable pour garantir que l’ACP ne privilégie pas certaines variables.
De manière générale, **la normalisation permet de mettre toutes les variables sur la même échelle**, ce qui est essentiel pour des modèles comme **KNN**, **K-means** ou encore pour l’ACP elle-même.

Résultat : mes modèles sont plus équilibrés, plus rapides à converger et donnent de meilleures performances.

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 !