**Faculté des sciences économiques et de gestion de Nabeul**

****

**Mastère professionnel en INGENIRIE ECONOMIQUE ET FINANCIERE**

**Année universitaire : 2023/2024**

**[Informatique Décisionnelle](https://www.facebook.com/groups/501804388697485/)**

**Encadrer par :**

**Mme Fadoua Bouafif**

**Réalise par : Ons belaid**

**Ameni ben othmene**

**Sommaire**

**Introduction**

**Chapitre1 : Recherche et collecte des données**

**Chapitre 2 : Exploration des données**

**Chapitre 3 : La relation entre l’âge et le balance des soldes bancaires « nuages des points ».**

**Chapitre 4 : La relation de balance par rapport au duration « courbe ».**

**Chapitre 5 : La relation de l’âge et le balance « histogramme ».**

**Introduction**

**Présentation de l'ensemble de données**

* Il existe différentes fonctionnalités (colonnes) dans nos données. Ces fonctionnalités incluent des informations démographiques et économiques sur une personne. L'analyse de ces données peut être effectuée à l'aide de diverses techniques d'apprentissage automatique ou d'analyse de données.
* Par exemple, des modèles de classification peuvent être formés sur ces données. Les données indiquent si une demande de prêt a été approuvée par les personnes répertoriées dans la colonne « approbation ». Les données des colonnes « âge », « emploi », « situation conjugale », « scolarité », « solde », « logement », « durée » et « campagne » décrivent les caractéristiques de ces personnes. Les effets de ces fonctionnalités sur l'approbation du prêt peuvent être explorés et un modèle d'apprentissage automatique peut être formé pour prédire si les demandes de prêt seront approuvées.
* Les relations entre les données peuvent également être examinées. Il est par exemple possible de vérifier s'il existe une corrélation entre « solde » et « durée ». Cela peut aider à comprendre s'il existe une relation entre le solde du compte bancaire d'une personne et la durée de remboursement du prêt.
* En conséquence, diverses analyses peuvent être faites sur les données dont nous disposons. À l’aide de techniques d’apprentissage automatique, l’effet des caractéristiques des données sur l’approbation des demandes de prêt peut être étudié et des modèles de prédiction peuvent être construits. Il permet également d'obtenir différents points de vue sur les relations et les caractéristiques entre les données, et ces informations peuvent aider à développer un système d'évaluation des demandes de prêt.

**Objectif de projet :**

Les données indiquent si une demande de prêt a été approuvée par les personnes répertoriées

**Chapitre 1 : recherche et collecte des données**

La collecte de données est une étape importante pour tout projet de recherche visant à étudier l'impact des différentes variables telque le solde bancaire « balance », la situation familiale, l’emploi sur l’approbation des prêts.

Les étapes de la collecte des données :

**1. Sélection des Sources de Données :** ( [Financial data of 4400+ public companies (kaggle.com)](https://www.kaggle.com/datasets/qks1lver/financial-data-of-4400-public-companies) )

**2.Identification des variables :**

* Age
* Emploi
* Matrimonial ou la situation conjugale ou familiale
* Education ou scolarité
* Balance ou le solde bancaire
* Logement c’est à dire que si les personnes ont des crédits de logement
* Durée de remboursement des crédits accordées
* Campagne
* Approbation

**3. Collecte des Données** : collecter les informations sur ces variables pertinentes

**Chapitre 2 : Exploration des données**

Pour l'exploration des données en Python, plusieurs bibliothèques sont largement utilisées. Python, avec ses bibliothèques riches et puissantes telles que Pandas, Numpy, Matplotlib et Seaborn, offre un cadre flexible et efficace pour explorer, nettoyer, visualiser et analyser des données de manière approfondie.

**Manipulation des données**

**Pandas** : est une bibliothèque Python puissante et populaire pour la manipulation et l'analyse de données. Elle fournit des structures de données flexibles et performantes, ainsi que des outils pour effectuer des opérations de manipulation et d'analyse de données.

**NumPy** est une bibliothèque fondamentale pour le calcul numérique en Python. Elle offre des tableaux multidimensionnels et des fonctions pour effectuer des opérations mathématiques sur ces tableaux.

**Statistics :** offre des fonctions de base pour calculer des statistiques descriptives. Cependant, pour des tâches statistiques plus avancées et des analyses exploratoires de données, les bibliothèques telles que NumPy, SciPy et Pandas sont souvent plus utilisées.

Python, avec ses bibliothèques riches et puissantes telles que Pandas, Numpy, Matplotlib et Seaborn, offre un cadre flexible et efficace pour explorer, nettoyer, visualiser et analyser des données de manière approfondie.

**Visualisation des données :**

Pour la visualisation des informations sur les données en Python, plusieurs bibliothèques sont couramment utilisées. Les trois principales bibliothèques de visualisation de données sont :

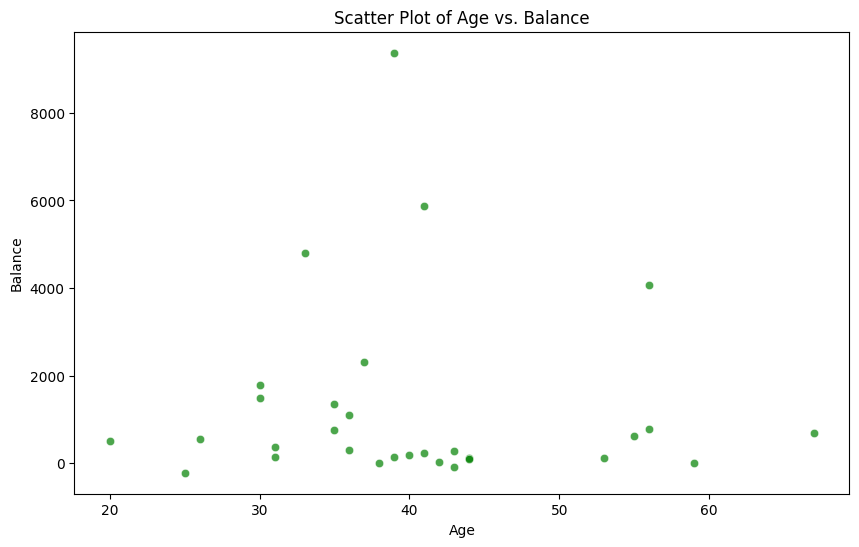
**Matplotlib**: c’est une bibliothèque de visualisation de données en Python qui offre une grande flexibilité pour créer une grande variété de graphiques et de visualisations. Matplotlib est souvent utilisé conjointement avec des bibliothèques comme NumPy et Pandas pour analyser et visualiser des données de manière efficace.

**Seaborn**: c’est une bibliothèque de visualisation de données en Python qui est basée sur Matplotlib. Elle offre une interface de haut niveau pour créer des graphiques statistiques attractifs et informatifs. Seaborn est particulièrement utile pour explorer et visualiser des relations dans des ensembles de données complexes.

**Chapitre 3 : la relation entre l’âge et le balance des comptes bancaires**

Les diagrammes de dispersion sont utiles pour visualiser la distribution des données et d’identifier des schémas potentiels ou des valeurs aberrantes. Ils peuvent fournir des informations sur la relation entre deux variables.

* Axe-X (horizontale) : Représente l'âge des individus.
* Axes-Y (verticale) : Représente le solde des comptes bancaires.



En examinant la dispersion des points, on peut rechercher des tendances ou des motifs visuels dans la relation entre l'âge et le solde bancaire.

Par exemple, une concentration de points dans une région particulière peut indiquer une corrélation entre un groupe d'âge spécifique et des soldes des comptes.

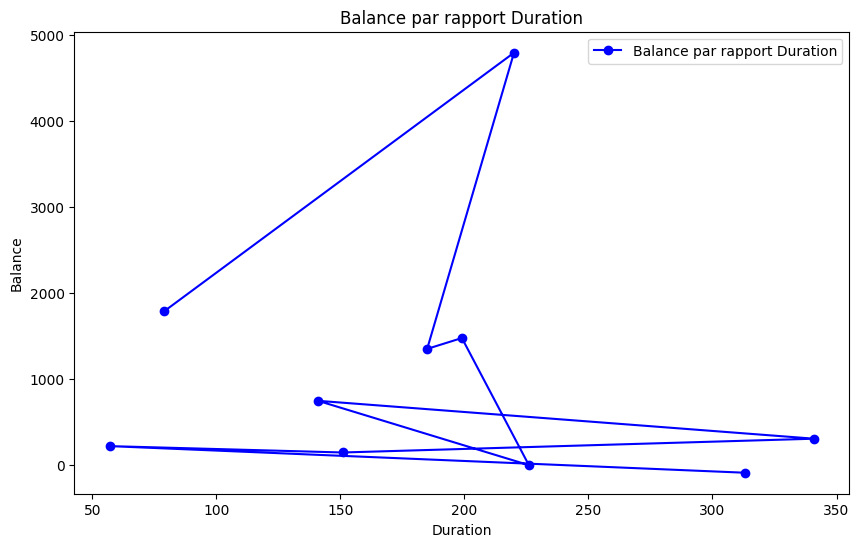
🡺 On remarque qu’Il y a une concentration de distribution des points ou les personnes âgées entre 30 et 45 ans ont des soldes bancaires « balance » entre 0 et 2000 euros.

**Chapitre 4 : la relation de Balance par rapport Duration**

Le graphique représente la relation entre la durée « duration » sur l’axe des X et le solde bancaire « balance » sur l’axe des Y. La courbe montre comment le solde varie en fonction de la durée.

Chaque point représente une observation d’individu et la ligne reliant les points peut aider à identifier des tendances.

La légende indique que la ligne représente la relation "Balance par rapport Duration"



Puisque les valeurs sont aléatoires, la courbe n’est pas trop significative mais on constate que plus les soldes bancaires sont faibles plus la durée de remboursement de prêts est plus longue.

**Chapitre 5 la relation de l’âge et le balance**

L'histogramme représente la distribution de l'âge en fonction des soldes bancaires ;

* L'axe des x représente les différentes tranches d'âge, regroupées en fonction du nombre spécifié de bacs.
* L'axe des y représente la fréquence (nombre d'occurrences) dans chaque tranche d'âge.

La forme de l'histogramme peut fournir des informations sur la répartition de l'âge, par exemple, il y a une concentration dans une plage d'âge pour les personnes de 25 ans jusqu’à 50ans ont les soldes bancaires élevé supérieur à 3000 euros.

