Rapport d'Analyse des Préférences d'Investissement

BRAKOUK

FATIMA

3 décembre 2025

1 Introduction

Ce rapport transforme le notebook Jupyter "CC-1.ipynb" en un document structuré. Il analyse un dataset de 40 répondants sur leurs préférences d'investissement, chargé depuis des fichiers CSV (AA.csv et BB.csv) via Google Colab, incluant des importations de bibliothèques comme Pandas, Seaborn et Matplotlib, ainsi qu'un téléchargement Kaggle d'un dataset financier.[1]

2 Description du Dataset

Le dataset "Financedata" comporte 24 colonnes et 40 observations, avec des variables numériques (âge moyen 27,8 ans, écarts-types sur préférences d'investissement de 1,13 à 1,80) et catégorielles (genre, durée d'investissement, fréquence de monitoring, attentes de rendement).

Les préférences de classement (Mutual Funds, Equity Market, etc.) varient de 1 (préféré) à 7, avec Mutual Funds en tête (moyenne 2,55) et Gold en bas (5,98). Les répondants investissent majoritairement dans Mutual Funds ou Equity, pour des objectifs comme la retraite ou l'appréciation du capital.[1]

3 Statistiques Descriptives

Variable	Moyenne	Écart-type	Min	Médiane	Max
Âge	27,80	3,56	21	27	35
Mutual Funds	2,55	1,20	1	2	7
Equity Market	3,48	1,13	1	4	6
Debentures	5,75	1,68	1	6,5	7
Government Bonds	4,65	1,37	1	5	7
Fixed Deposits	3,58	1,80	1	3,5	7
PPF	2,03	1,61	1	1	6
Gold	5,98	1,14	2	6	7

TABLE 1 – Statistiques descriptives des variables numériques

Toutes les colonnes sont non nulles, avec une consommation mémoire de 40 KB.[1]

4 Visualisations et Insights

Le notebook génère des histogrammes pour les distributions de préférences (Mutual Funds à Gold), montrant une préférence marquée pour les fonds mutuels et PPF, et un rejet des obligations et or. Des graphiques comme ceux de Seaborn illustrent les tendances par âge et genre, avec une attente moyenne de rendement de 20-30 %. Les sources d'information incluent journaux, consultants et internet.[1]

Ce format conserve l'essence analytique du notebook tout en le rendant lisible comme un rapport académique ou professionnel.[1]

5 Interprétation du code et des résultats du notebook Jupyter

Le notebook effectue une analyse exploratoire de données (EDA) sur un jeu de données financier (apparemment des données d'enquête sur l'investissement) et tente ensuite de construire des modèles d'apprentissage automatique (machine learning) pour prédire des variables cibles.

5.1 Préparation et chargement des données

Cellule [1]

Action : Téléchargement du jeu de données Kaggle nitindatta/finance-data.
Interprétation : Le jeu de données est téléchargé et extrait avec succès. Le chemin d'accès aux fichiers est affiché.

Cellule [2]

Action : Importation des bibliothèques Python nécessaires (numpy, pandas, seaborn, matplotlib, glob, os, warnings).
Interprétation : Préparation de l'environnement pour l'analyse et la visualisation des données.

Cellule [5]

Action : Listing des fichiers (AA.csv, BB.csv) dans le répertoire d'entrée.
Interprétation : Confirme la présence des deux fichiers CSV qui seront utilisés.

Cellule [7]

Action : Chargement de AA.csv dans Original_data et de BB.csv dans Finance_data (DataFrames pandas).
Interprétation : Les deux jeux de données sont chargés en mémoire.

Cellule [8]

Action : Affichage des 5 premières lignes de Original_data.
Interprétation : Montre les colonnes brutes de l'enquête, avec des noms très longs (par

ex. *What do you think are the best options for investing your money ? (Rank in order of preference) [Mutual Funds]*).

Cellules [9]

Action : Affichage des 5 premières et 5 dernières lignes de Finance_data.
Interprétation : Finance_data est une version nettoyée de Original_data, avec des noms de colonnes plus courts (par ex. Mutual_Funds, Equity_Market, gender, age). C'est ce DataFrame qui sera utilisé pour l'analyse et le ML.

Cellules [10], [11]

Action : Vérification des valeurs manquantes (isnull().sum()).
Interprétation : Aucune valeur manquante n'est détectée dans Original_data ni dans Finance_data.

Cellule [12]

Action : Affichage des informations des DataFrames (info()).
Interprétation : Chaque DataFrame a 40 entrées et 24 colonnes, dont 8 colonnes int64 (classements d'investissement) et 16 colonnes object (variables catégorielles).

Cellules [13], [14]

Action : Vérification des lignes dupliquées et de la taille (shape).
Interprétation : Aucune ligne entièrement dupliquée. Les deux DataFrames ont la même taille (40 lignes, 24 colonnes).

Cellule [15]

Action : Comptage des valeurs uniques par colonne (nunique()).
Interprétation : Le jeu de données est petit (40 observations). Les colonnes numériques ont 6 à 7 valeurs uniques, les catégorielles 2 à 4.

Cellule [16]

Action : Affichage des types de données (dtypes).
Interprétation : Confirme la structure : 8 colonnes int64, 16 colonnes object.

Cellules [17], [18]

Action : Séparation des colonnes en types object (category_type) et numériques (Number_type).
Interprétation : Prépare l'analyse en fonction du type de variable.

Cellule [19]

Action : Informations sur l'utilisation mémoire (info(memory_usage='deep')).
Interprétation : Chaque DataFrame utilise environ 40 KB, confirmant la petite taille.

Cellule [20]

Action : Statistiques descriptives pour Finance_data (describe().T).
Interprétation :

- L'âge varie de 21 à 35 ans (moyenne 27,8).
- Les classements (1 = meilleur, 7 = pire) montrent que les fonds mutuels (Mutual_Funds) et le PPF (PPF) ont les moyennes les plus faibles (options préférées).
- Les débentures (Debentures) et l'or (Gold) ont les moyennes les plus élevées (options moins préférées).

Cellules [21], [22]

Action : Listage des colonnes numériques et catégorielles.
Interprétation : Confirme les colonnes de chaque type.

5.2 Visualisation et analyse exploratoire (EDA)

Cellule [23]

Action : Histogrammes des distributions de classements d'investissement.
Interprétation :

- PPF et Mutual_Funds : classements plutôt bas (1–3), options privilégiées.
- Debentures et Gold : classements élevés (6–7), options peu appréciées.
- Equity_Market et Fixed_Deposits : distributions plus équilibrées ou bimodales.

Cellule [24]

Action : Diagramme à barres du taux d'investissement par genre (gender).
Interprétation : Répartition relativement équilibrée entre femmes et hommes (environ 20 chacun).

Cellule [25]

Action : Diagramme à barres du nombre de participants par groupe d'âge et genre.
Interprétation : Le groupe 23–40 ans domine. Le groupe 41+ ans n'apparaît pas.

Cellule [26]

Action : Diagramme à barres de la durée d'investissement (Duration).
Interprétation : Les durées *1–3 years* et *3–5 years* sont majoritaires. *Less than 1 year* est rare.

Cellule [27]

Action : Diagramme à barres des facteurs influençant l'investissement (Factor).
Interprétation : Le risque (*Risk*) est le facteur le plus cité, suivi des rendements (*Returns*) ; la sécurité (*Safety*) est la moins citée.

Cellule [28]

Action : Diagramme à barres des investisseurs masculins par âge (filtré par Investment_Avenues == 'Yes').
Interprétation : Les hommes dans la vingtaine et la trentaine sont les plus nombreux (pics à 24, 27, 30 ans).

Cellule [29]

Action : Graphique linéaire des investisseuses féminines par âge (même filtre).
Interprétation : Pics à 23, 24, 30 et 34 ans, distribution proche de celle des hommes.

Cellule [30]

Action : Diagramme circulaire des sources d'information (Source).
Interprétation :

- *Financial Consultants* : 35,1 %.
- *Internet* : 32,4 %.
- *Newspapers and Magazines* et *Television* : moins utilisées.

Cellules [31], [32]

Action : Matrice de corrélation et carte de chaleur (variables numériques).
Interprétation :

- Corrélation faible et positive entre l'âge et Equity_Market (0,25) et Debentures (0,33).
- Corrélation négative notable : Debentures vs PPF (-0,51) et Government_Bonds vs Fixed_Deposits (-0,53).

Cellules [33], [34]

Action : Comptage et visualisation des investissements pour Source == 'Financial Consultants'.
Interprétation : Debentures et Gold ont les totaux de classement les plus élevés (97 chacun, donc moins préférés). Le PPF (25) est le plus apprécié.

5.3 Modélisation par apprentissage automatique

Prédiction de la variable Stock_Marktet

Problème : classification binaire (Yes/No).
Taille du test : 8 observations (2 de classe 0, 6 de classe 1).
Résultats :

Modèle	Accuracy	F1-Score (classe 0)	F1-Score (classe 1)
Logistic Regression	0,75	0,00	0,86
Random Forest	0,75	0,00	0,86
SVM	0,75	0,00	0,86

Interprétation :

- L'accuracy de 0,75 est trompeuse : la classe 0 est totalement ignorée.

— Le déséquilibre des classes (peu de 0) pousse les modèles à prédire uniquement la classe 1.
— Conclusion : les modèles ne savent pas distinguer la classe 0, les résultats ne sont pas fiables.

Prédiction de la variable Duration

Problème : classification multi-classes (4 durées possibles).
Résultats simples :

Modèle	Accuracy	Moyenne CV Accuracy
Logistic Regression	0,625	0,675
Random Forest	0,875	0,650
SVM	0,375	0,350

Résultats GridSearch (CV) :
— Logistic Regression : meilleur score CV 0,619.
— SVM : meilleur score CV 0,590.
— Random Forest : meilleur score CV 0,595.
Évaluation finale du Random Forest optimisé :
— Accuracy sur le jeu de test : 0,875.
Interprétation :
— Le Random Forest est le meilleur modèle pour prédire Duration sur ces données.
— Les avertissements de convergence pour la régression logistique et la faible performance du SVM suggèrent une sensibilité à la petite taille de l'échantillon et à la normalisation.

Références

[1] Analyse du fichier CC-1.ipynb (725 KB), 2025.