

# Rapport Mondial sur le Bonheur 2016

Analyse de Données et Modélisation Prédictive

Ayoub Enouijai

3 décembre 2025

## Résumé

Ce rapport présente une analyse complète du rapport mondial sur le bonheur 2016. Nous explorons l'ensemble de données pour identifier les principaux moteurs du bonheur à travers 157 pays et appliquons des algorithmes d'apprentissage automatique (Régression Linéaire, Forêt Aléatoire et SVR) pour prédire les scores de bonheur. Le Régresseur par Forêt Aléatoire a donné les meilleures performances avec un score  $R^2$  de 0.8194.

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Méthodologie et Préparation des Données</b>	<b>2</b>
2.1	Chargement des Données . . . . .	2
2.2	Sélection des Variables et Normalisation . . . . .	2
<b>3</b>	<b>Analyse Exploratoire des Données (AED)</b>	<b>2</b>
3.1	Analyse de Corrélation . . . . .	2
3.2	Visualisation des Distributions . . . . .	3
<b>4</b>	<b>Modélisation Prédictive</b>	<b>4</b>
4.1	Comparaison des Modèles . . . . .	4
4.2	Optimisation et Évaluation Finale . . . . .	5
<b>5</b>	<b>Conclusion</b>	<b>5</b>

# 1 Introduction

Le Rapport Mondial sur le Bonheur est une enquête repère sur l'état du bonheur global. Notre objectif est de comprendre la corrélation entre le score de bonheur et divers facteurs socio-économiques comme le PIB par habitant, le soutien familial et l'espérance de vie en bonne santé, puis de construire des modèles pour prédire ce score.

## 2 Méthodologie et Préparation des Données

### 2.1 Chargement des Données

L'ensemble de données `2016.csv` est chargé. L'inspection des données montre qu'il contient 157 lignes (pays) et 13 colonnes (variables).

```
1 nRowsRead = 1000
2 df2 = pd.read_csv('2016.csv', delimiter=',', nrows = nRowsRead)
3 df2.dataframeName = '2016.csv'
4 nRow, nCol = df2.shape
5 print(f'Il y a {nRow} lignes et {nCol} colonnes')
6 # Sortie: Il y a 157 lignes et 13 colonnes
```

Listing 1 – Chargement du jeu de données

### 2.2 Sélection des Variables et Normalisation

Le Score de Bonheur est la variable cible. Les facteurs retenus (PIB, Famille, Santé, etc.) sont normalisés à l'aide de `StandardScaler` pour éviter que des échelles différentes n'influencent les modèles, avant d'être divisés en ensembles d'entraînement et de test.

## 3 Analyse Exploratoire des Données (AED)

### 3.1 Analyse de Corrélation

La matrice de corrélation, présentée ci-dessous, met en évidence que le **PIB par habitant** et la **Santé** sont les facteurs les plus fortement corrélés positivement avec le Score de Bonheur.

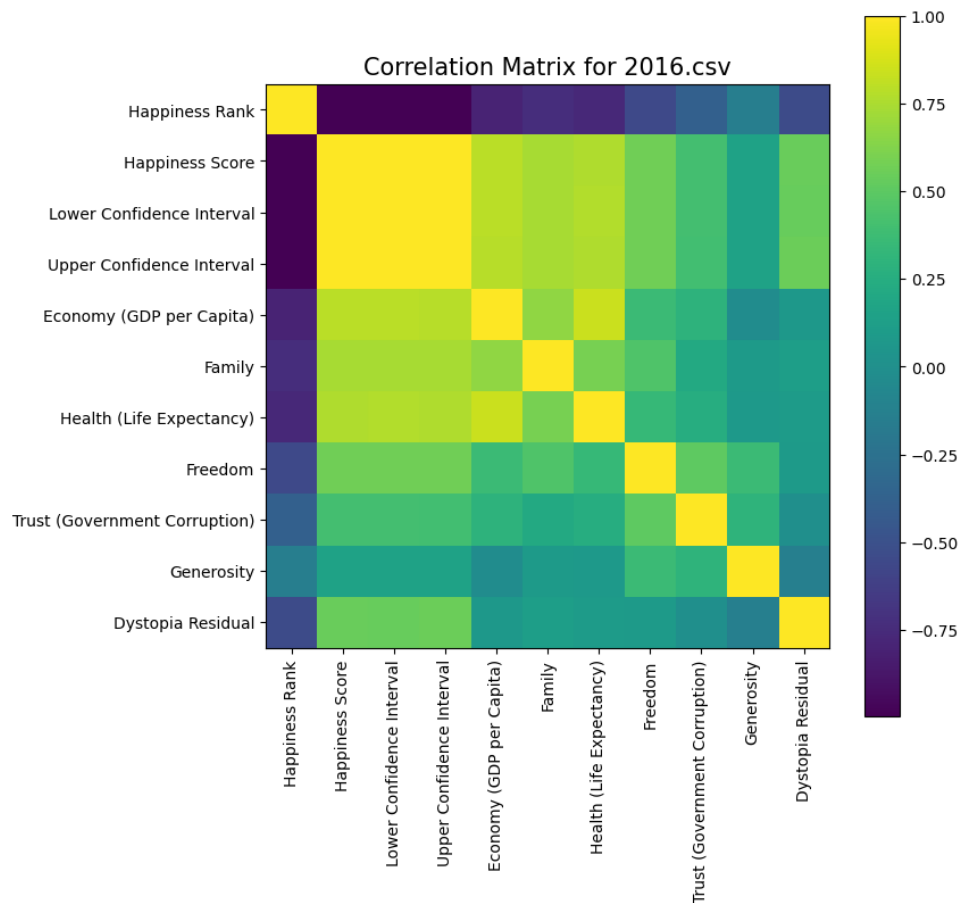


FIGURE 1 – Carte de chaleur des corrélations entre les métriques de bonheur. Une forte corrélation est observée entre le Score de Bonheur et l'Économie.

### 3.2 Visualisation des Distributions

Les relations par paire des variables ont été examinées pour détecter des tendances. La figure suivante présente un aperçu de ces relations, tel que généré par la fonction `plotScatterMatrix` de notre notebook.

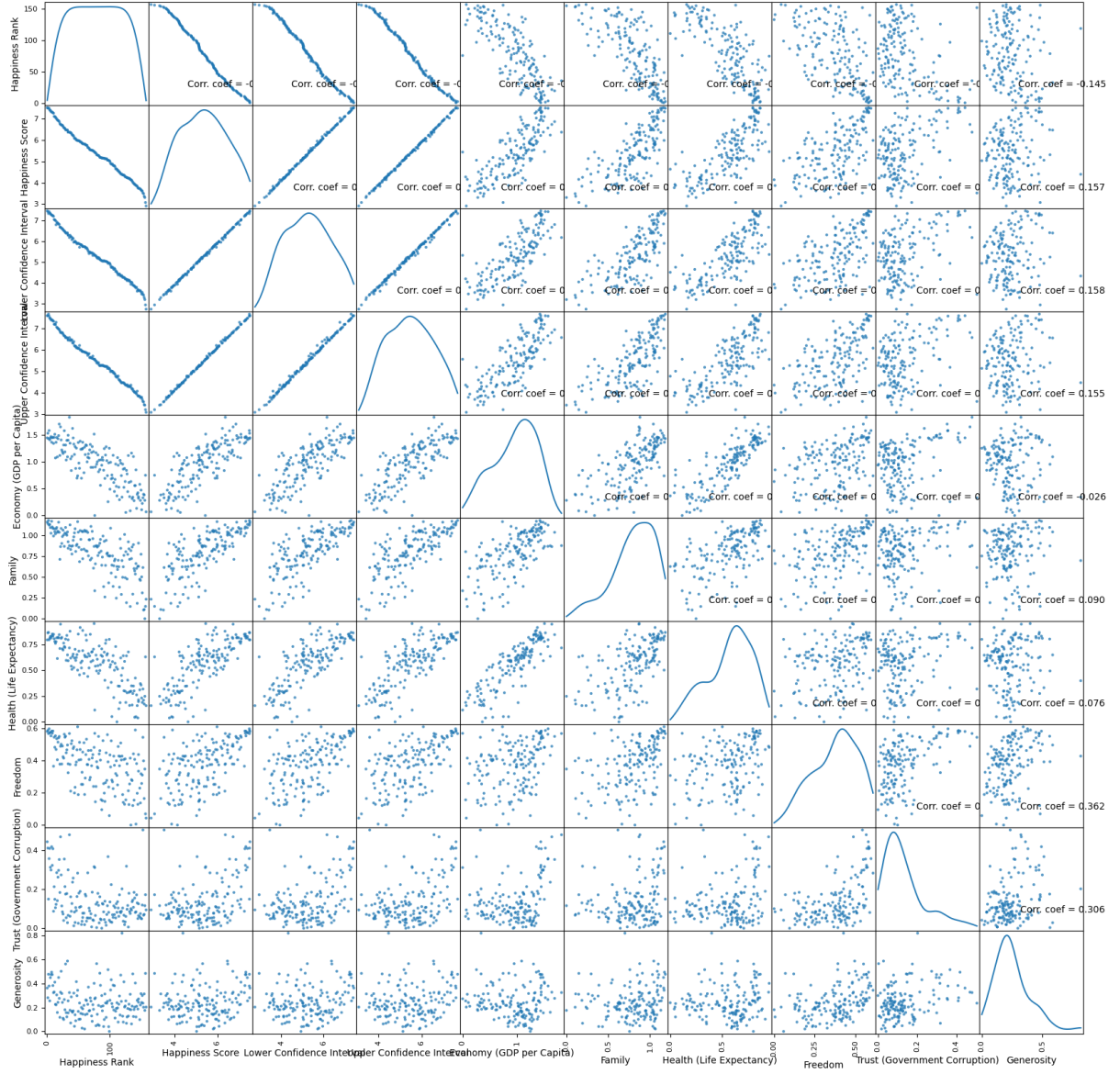


FIGURE 2 – Visualisation des distributions et des relations par paires de certaines caractéristiques clés.

## 4 Modélisation Prédictive

### 4.1 Comparaison des Modèles

Trois modèles de régression ont été évalués par validation croisée (5 plis). Les performances initiales sont présentées dans le Tableau 1.

TABLE 1 – Performance des modèles par Validation Croisée (Score  $R^2$ )

Modèle	Score $R^2$ Moyen
Régression Linéaire	0.7364
Forêt Aléatoire (Random Forest)	0.7328
SVR (Régresseur à Vecteur de Support)	0.7322

## 4.2 Optimisation et Évaluation Finale

Le modèle de **Forêt Aléatoire** a été sélectionné pour l'optimisation. L'ajustement des hyperparamètres par **GridSearchCV** a porté le score  $R^2$  moyen de validation à 0.7442.

L'évaluation finale sur l'ensemble de test a produit les résultats suivants :

- **RMSE (Erreur Quadratique Moyenne) :** 0.4844
- **Score  $R^2$  :** 0.8194

Le score  $R^2$  final de 0.8194 indique que le modèle de Forêt Aléatoire explique environ 82% de la variance dans le Score de Bonheur.

## 5 Conclusion

L'analyse a confirmé que les facteurs socio-économiques sont des prédictors robustes du bonheur national. Le modèle de **Forêt Aléatoire** optimisé est le plus performant, capturant efficacement les relations complexes et non linéaires entre le bonheur et ses composantes, et constituant une excellente base pour une prédiction fiable.