**Project 2024**

Introduction to Machine Learning

**Group Name**: The Happiness Duo

Student # 1:

* **Full Name**: LAURENT Sacha
* **Student Number/ID**: 20220702
* **Work done**: Linear regression, K-Means clustering, PCA, Error metrics

Student # 2:

* **Full Name**: SIMON Eliot
* **Student Number/ID**: 20220758
* **Work done:** KNN classifier, HCA, PCA, Decision tree classifier

Table of Contents

[Step 1: Dataset Selection 4](#_Toc184131853)

[Step 2: Scenario/About Dataset: 4](#_Toc184131854)

[Problem Statement: 4](#_Toc184131855)

[How Machine Learning is Useful on this dataset: 4](#_Toc184131856)

[Step 3: Data Loading 5](#_Toc184131857)

[Step 4: Data Wrangling or Data Pre-processing 5](#_Toc184131858)

[Handle missing values 5](#_Toc184131859)

[Standardize the data 6](#_Toc184131860)

[Step 5: Exploratory Data Analysis 6](#_Toc184131861)

[Happiness Score Histogram 6](#_Toc184131862)

[Happiness Score vs Economy (GDP per Capita) 7](#_Toc184131863)

[Happiness Score vs Health (Life Expectancy) 8](#_Toc184131864)

[Happiness Score vs Generosity 8](#_Toc184131865)

[Step 6: Model Development 9](#_Toc184131866)

[KNN Classification 9](#_Toc184131867)

[Define the classes 9](#_Toc184131868)

[Split the data 9](#_Toc184131869)

[Fit the model and classify the test data 9](#_Toc184131870)

[Linear Regression 10](#_Toc184131871)

[Selecting features 10](#_Toc184131872)

[Split the data 10](#_Toc184131873)

[Fit and Obtain the coefficients of the model 10](#_Toc184131874)

[Plot the curve 11](#_Toc184131875)

[K-Means clustering 12](#_Toc184131876)

[Selecting features 12](#_Toc184131877)

[Fit the data 12](#_Toc184131878)

[Plot the clusters 12](#_Toc184131879)

[HCA 14](#_Toc184131880)

[Selecting features 14](#_Toc184131881)

[Dendrogram 14](#_Toc184131882)

[Plot the clusters 15](#_Toc184131883)

[PCA 16](#_Toc184131884)

[Correlation Matrix 16](#_Toc184131885)

[Determine the number of components (new features) we will use 17](#_Toc184131886)

[Features / Component Correlation 17](#_Toc184131887)

[Correlation circle 18](#_Toc184131888)

[Final plot 19](#_Toc184131889)

[Decision Tree Classifier on the PCA’s data 20](#_Toc184131890)

[Step 7: Model Evaluation 22](#_Toc184131891)

[KNN Classification 22](#_Toc184131892)

[Accuracy Evaluation 22](#_Toc184131893)

[Linear Regression 23](#_Toc184131894)

[Regression Plot 23](#_Toc184131895)

[K-Means clustering 24](#_Toc184131896)

[Elbow method 24](#_Toc184131897)

[Silhouette score 25](#_Toc184131898)

[Step 8: Model Refinement 26](#_Toc184131899)

[Linear Regression 26](#_Toc184131900)

[K-Means clustering 27](#_Toc184131901)

[Comparison with HCA 28](#_Toc184131902)

[PCA 29](#_Toc184131903)

[Decision Tree Classifier on the PCA’s data 30](#_Toc184131904)

# Step 1: Dataset Selection

We chose the **World Happiness Report** dataset.

It has features such as Happiness scored according to economic production, social support, etc.

<https://www.kaggle.com/datasets/unsdsn/world-happiness> Dataset link on Kaggle

# Step 2: Scenario/About Dataset:

This dataset contains global data on happiness scores and factors that contribute to happiness, such as GDP per capita (Economy), family, social support, life expectancy ...

## Problem Statement:

* We want to understand global happiness trends to design effective policy recommendations and allocate resources for improving well-being across different regions. With the World Happiness dataset, which includes metrics such as GDP per capita, social support, life expectancy, freedom, corruption ... Machine Learning will allow us to identify key trends and actionable insights

## How Machine Learning is Useful on this dataset:

* **Regression**: To predict happiness scores based on measurable socio-economic indicators, providing a model able to predict happiness scores and simulate the impact of policy changes.
* **Clustering**: To group countries into similar categories based on happiness profiles or socioeconomic factors, revealing hidden patterns regional similarities and global trends
* **Classification**: To categorize countries into predefined happiness levels (e.g., "High Happiness," "Medium Happiness," "Low Happiness") for prioritizing regions needing immediate attention

# Step 3: Data Loading

Une image contenant texte, capture d’écran, conception

Description générée automatiquementHere is a short view of the dataset after loading it into python.

We can see all the features and have an overview of the values for the different countries.

# Step 4: Data Wrangling or Data Pre-processing

## Handle missing values

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

With a simple python script, we can see that many values are missing for each column, we will handle them by filling them with the **mean of the column**

## Standardize the data

Now that no more values are missing, we can **Standardize the data** after Dropping the useless columns (String values such as “Country”, “Region” and global descending rank “Happiness Rank”)

Une image contenant texte, capture d’écran, Police, menu

Description générée automatiquement

# Step 5: Exploratory Data Analysis

In order to understand better the dataset and identify relevant features, we will apply various data exploration techniques.

## Happiness Score Histogram

Une image contenant texte, capture d’écran, diagramme, Police

Description générée automatiquement

This histogram reveals the average World happiness. We can see many countries manage to have a Happiness score above 7/10 although **the most frequent values are between 4 and 5 /10**.

A first intuition would be to think that **many variables are linearly linked with the Happiness Score**.  
Let’s see if this intuition is correct.

## Happiness Score vs Economy (GDP per Capita)

Une image contenant texte, capture d’écran

Description générée automatiquement

This graph seems to show that the more money a country produces, the happier its inhabitants will be.

Une image contenant texte, capture d’écran, Police, Caractère coloré

Description générée automatiquementUne image contenant capture d’écran, texte, Caractère coloré

Description générée automatiquement

Here is the same graph as above but colored by which region the countries are in.

We can see that **the countries in the same region tend to have similar happiness and economy scores,** which is a very important insight for our future analysis.

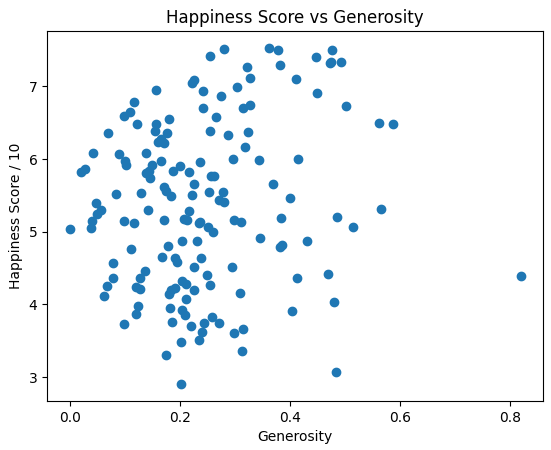
## Happiness Score vs Health (Life Expectancy)

Une image contenant capture d’écran, texte

Description générée automatiquement

This graph seems to further confirm that many variables are linearly linked with the Happiness Score as the better the health a country has, the happier its inhabitants will be.

## Happiness Score vs Generosity



However, it is important to note that not all features behave that way. **Happiness does not seem to depend linearly on the Generosity for example** which was against our first intuition.

# Step 6: Model Development

## KNN Classification

### Define the classes

First, we will retrieve the Happiness score (grade/10) of all countries and use these values as the classes for the KNN classifier.

Une image contenant texte, Police, capture d’écran, typographie

Description générée automatiquement

### Split the data

Then, to test the KNN algorithms, we will split the data into training and testing sets using **20%** of the dataset for testing.

### Fit the model and classify the test data

Using sklearn.neighbors KNeighborsClassifier we can easily fit the model to our dataset and see that the model is indeed classifying the countries of the test set into categories (happiness score/10).

Here are the classes found for our test data:



We will test the accuracy of the model in the Model Evaluation step and interpret the results in the Model Refinement step

## Linear Regression

### Selecting features

For this Linear Regression, we will choose "Economy (GDP per Capita)" as our independent variable and "Happiness Score" as our dependent variable.

Here we recall the graph of the Happiness Score in function of the Economy (GDP per Capita)

Une image contenant texte, capture d’écran

Description générée automatiquement

### Split the data

Then, to test our algorithm, we will split our data into training and testing sets using **20%** of the dataset for testing.

### Fit and Obtain the coefficients of the model

Using sklearn.linear\_model LinearRegression, we can easily implement our Linear Regression algorithm and fit our data.

Then, we obtain the coefficients of the model and are able to plot the curve of equation

### Plot the curve

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

This model clearly fits the data really well. The Happiness score and the Economy (GDP per Capita) are very linearly dependent.

We will test the accuracy of the model in the Model Evaluation step and interpret the results in the Model Refinement step

## K-Means clustering

### Selecting features

For this K-Means clustering, we will choose "Economy (GDP per Capita)" as our independent variable and " Generosity" as our dependent variable.

These variables are really interesting because, unlike what we might think, they are not linearly dependent. This Clustering should allow us to understand the moral values of the countries based on their economy.

Here we recall the graph of the Happiness Score in function of the Economy (GDP per Capita)

Une image contenant capture d’écran, texte, diagramme

Description générée automatiquement

### Fit the data

Using sklearn.cluster KMeans we can easily implement our K-Means clustering algorithm and fit our data.

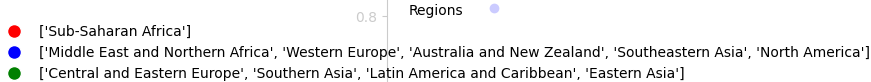
### Plot the clusters

We can then plot the clusters using our 2 variables.

To really understand the meaning of the clusters, we also plotted the un-clustered graph but with different colors for regions (manual classifying). The interpretation of the results is in in the Model Refinement step

Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement

Comparison with the graph colored by regions (see graph legend):Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement Graph legend:

## HCA

### Selecting features

Let’s choose the same variables as the K-Means clustering. These variables are relevant for clustering and we will be able to compare the 2 models.

### Dendrogram

Using scipy.cluster.hierarchy we can easily create a Dendrogram representation of HCA.

Une image contenant texte, diagramme, Plan, Rectangle

Description générée automatiquement

We can see that 3 seems to be a very fitting number of clusters, similarly to K-Means.

### Plot the clusters

Then, using sklearn.cluster AgglomerativeClustering we can implement our HCA clustering algorithm and fit our data.

Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement

The clusters are better defined since the model is not sensitive to noise and are less spherical. However, as a consequence, they also correspond less to the classes we found in the graph colored by region above, making these clusters a bit **meaningless to our data**. The interpretation of the results is in in the Model Refinement step.

## PCA

PCA is a dimensionality reduction technique to **simplify datasets** with many features while retaining as much variability (information) as possible. It works by:

* Identifying the directions (principal components) along which the data varies the most (most “powerful” features).
* Transforming the original features into a new set of uncorrelated features (principal components) ordered by the amount of variance they capture.
* Reducing the dataset to **fewer dimensions** by selecting the top principal components

### Correlation Matrix

The correlation matrix is the table showing the pairwise correlation coefficients between all variables in our dataset. Each cell in the matrix contains the correlation value (ranging from -1 to 1) for two variables:

* 1: Perfect positive correlation.
* 0: No correlation.
* -1: Perfect negative correlation.

We need it to assess relationships between the variables (prove that **there is redundancy** 🡪 PCA is relevant)

Une image contenant texte, capture d’écran, carré, Parallèle

Description générée automatiquement

We consider strong correlations values above 0.5 or below -0.5 (the red on the graph)

We can See That there is strong correlation between multiple variables, therefore The PCA is relevant to reduce the dimensionality, and the computation power needed for a sensible analysis

### Determine the number of components (new features) we will use

By diagonalizing the correlation matrix, we can deduce how many components are needed.

Une image contenant texte, diagramme, capture d’écran, ligne

Description générée automatiquement

We can see that keeping 2 principal components for the analysis retains about 70% of the information of the original data<br>

This is a good compromise between dimensionality reduction (and computation power) and information loss

### Features / Component Correlation

We verify that our features are represented well by the components

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

We can see that the components are strongly correlated with the original features

### Correlation circle

This is a plot of our old features based on their correlation with the new Components that we defined.

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

From the correlation circle, we can understand what each axis represents:

Axis 1 : Material and Economic Factors

* Family, health, economy

Axis 2 : Social and Moral Factors

* Generosity, Trust, freedom,

### Final plot

Une image contenant texte, diagramme, écriture manuscrite, ligne

Description générée automatiquementNow we will plot again the countries and the features but using the 2 principal components as the x and y axis

With a high enough treshold, we can clearly see two groups of counties. The PC1 component is linked to the countries wealth & PC2 component is linked to the Social/Moral Factors

## Decision Tree Classifier on the PCA’s data

Based on the PCA results, we can try to group our results, in this case, we will try to find out, if we can separate the occidental countries, from the rest of the world

First, we have to find the best depth for the classifier, then we can create the tree

* At each step, the algorithm evaluates all features to determine the best one to split the data.
* The "best" feature is ch osen based on a criterion that measures the quality of the split (e.g., Gini impurity or Entropy in Information Gain).

Une image contenant texte, diagramme, Plan, Rectangle

Description générée automatiquement

Finally, we can plot again the PCA’s data using these newly defined classes.

Une image contenant texte, diagramme, Tracé, ligne

Description générée automatiquement

On the final graph, we can see that the algorithm,managed to separate the occidental countries from the rest

# Step 7: Model Evaluation

## KNN Classification

### Accuracy Evaluation

Une image contenant texte, diagramme, Tracé, ligne

Description générée automatiquement

This graph shows us the accuracy of the model for different numbers of neighbors.

The best Test set accuracy was 0.78125 with k = 5

We can conclude that the best number of neighbors is 5 as it gives the highest accuracy for the test set while keeping a high accuracy for the training set.

## Linear Regression

Recall of the results:  
Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

### Regression Plot

This plot will show a combination of a scattered data points (a <b>scatterplot</b>), as well as the fitted <b>linear regression</b> line going through the data. This will give us a reasonable estimate of the relationship between the two variables, the strength of the correlation, as well as the direction (positive or negative correlation)

Une image contenant capture d’écran, texte, ligne

Description générée automatiquement

## K-Means clustering

Recall of the results obtained: Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement

### Elbow method

Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquement

We see a sudden drop in inertia at k=3, then the curve becomes linear.

Therefore, we can assume that the optimal number of clusters is 3 as we have previously chosen.

### Silhouette score

Une image contenant texte, capture d’écran, Tracé, Police

Description générée automatiquement

The silhouette score for k=3 is 0.44, which is not the highest, but is more than acceptable.

# Step 8: Model Refinement

## Linear Regression

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

This model further confirms our first intuition that **the more money a country produces, the happier its inhabitants will be**.

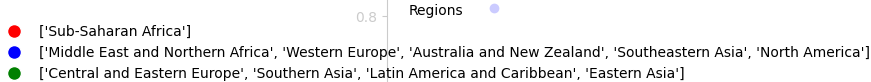
And since many other features are linearly dependent with the Happiness score, such as “Health (Life Expectancy)”, we can safely deduct the most important variable (those that have the higher impact on the happiness score)

## K-Means clustering

Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement

Graph colored by regions:Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement Graph legend:

With this graph, we can **explain better the clusters** generated by the K-Means algorithm:

Category 1 (red): "Emerging Regions"

* Includes: Sub-Saharan Africa
* Rationale: Sub-Saharan Africa is often characterized by developing economies and unique socio-political challenges, making it distinct from the other regions.

Category 2 (blue): "Developed and Diverse Economies"

* Includes: Middle East and Northern Africa, Western Europe, Australia and New Zealand, Southeastern Asia, North America
* Rationale: These regions represent a mix of developed economies and regions with significant economic diversity. They encompass high-income countries (e.g., Western Europe, North America, Australia and New Zealand) and rapidly developing areas with economic integration (e.g., Southeastern Asia, parts of the MiddleEast).

Category 3 (green): "Transitional and Emerging Economies"

* Includes: Central and Eastern Europe, Southern Asia, Latin America and Caribbean, Eastern Asia
* Rationale: These regions are characterized by a mix of developing and transitional economies, with many countries undergoing rapid industrialization and socio-economic changes.

## Comparison with HCA

Une image contenant capture d’écran, texte, Caractère coloré, diagramme

Description générée automatiquement

The model is not sensitive enough to sound to accurately cluster the countries/regions which is a **big limitation** in our case as the clusters are clear but meaningless regarding our data.

## PCA

Une image contenant texte, diagramme, écriture manuscrite, ligne

Description générée automatiquementRecall of the result:

PCA allowed us to:

* Identifying the directions (principal components) along which the data varies the most.
* Transforming the original features into a new set of uncorrelated features (principal components) ordered by the amount of variance they capture.
* Reducing the dataset to fewer dimensions by selecting the top principal components.

## Decision Tree Classifier on the PCA’s data

The Decision Tree Classifier provides a very insightful view

Advantages:

* Interpretable: The tree structure is intuitive and easy to understand.
* Non-linear: Can capture non-linear relationships between features and the target.
* Handles Mixed Data Types: Works with both numerical and categorical data.
* Requires Minimal Preprocessing: No need for feature scaling or centering.

However, when modifying parameters or updating the PCA, we realized the Decision Tree Classifier has many drawbacks:

* Overfitting: Decision trees can overfit the training data, capturing noise instead of the underlying pattern.
* Bias to Imbalanced Data: May perform poorly with imbalanced datasets unless balanced criteria are used.
* Instability: Small changes in data can lead to significantly different trees