**Project 2024**

Introduction to Machine Learning

**Group Name**: The Happiness Duo

Student # 1:

* **Full Name**: LAURENT Sacha
* **Student Number/ID**: 20220702
* **Work done**:

Student # 2:

* **Full Name**: SIMON Eliot
* **Student Number/ID**: 20220758
* **Work done:**

Table of Contents

[Step 1: Dataset Selection 3](#_Toc184121227)

[Step 2: Scenario/About Dataset: 3](#_Toc184121228)

[Problem Statement: 3](#_Toc184121229)

[How Machine Learning is Useful on this dataset: 3](#_Toc184121230)

[Step 3: Data Loading 4](#_Toc184121231)

[Step 4: Data Wrangling or Data Pre-processing 4](#_Toc184121232)

[Handle missing values 4](#_Toc184121233)

[Standardize the data 5](#_Toc184121234)

[Step 5: Exploratory Data Analysis 5](#_Toc184121235)

[Happiness Score Histogram 5](#_Toc184121236)

[Happiness Score vs Economy (GDP per Capita) 6](#_Toc184121237)

[Happiness Score vs Health (Life Expectancy) 7](#_Toc184121238)

[Happiness Score vs Generosity 7](#_Toc184121239)

[Step 6: Model Development 8](#_Toc184121240)

[KNN Classification 8](#_Toc184121241)

[Define the classes 8](#_Toc184121242)

[Split the data 8](#_Toc184121243)

[Fit the model and classify the test data 8](#_Toc184121244)

[Linear Reegression 9](#_Toc184121245)

[K-Means clustering 10](#_Toc184121246)

[PCA 11](#_Toc184121247)

[Decision Tree Classifier 12](#_Toc184121248)

# 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.

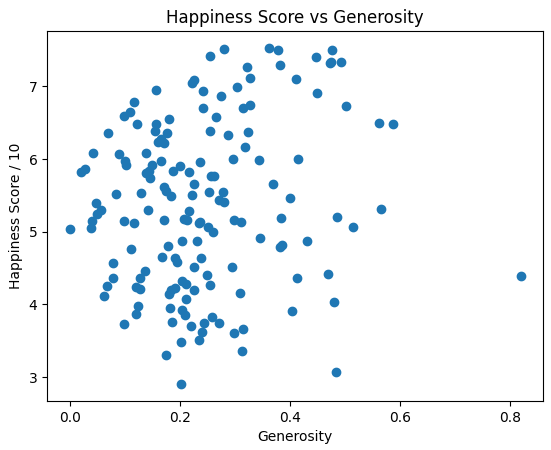
## Happiness Score vs Health (Life Expectancy)

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 Reegression

## K-Means clustering

## PCA

## Decision Tree Classifier