

## Workshop 3

Realizado por:  
Isaac Piedrahita

Docente  
Javier Alejandro Vergara Zorrilla  
ETL



Programa de Ingeniería de Datos e Inteligencia Artificial  
Facultad de ingeniería  
Cuarto Semestre  
Universidad Autónoma de Occidente  
Santiago de Cali  
2024

# Documentación del Proceso: Taller 3 - Machine Learning y Data Streaming

## Introducción

En este taller, trabajamos con cinco archivos CSV que contienen información sobre los niveles de felicidad en diferentes países durante varios años. El objetivo es entrenar un modelo de regresión para predecir la puntuación de felicidad, crear el flujo completo de EDA/ETL para extraer características de los archivos, entrenar el modelo usando una división de datos 70-30 (70% para entrenamiento y 30% para prueba), transmitir los datos transformados y, en el consumidor, usar el modelo entrenado para predecir la puntuación de felicidad y almacenar las predicciones en una base de datos junto con las características respectivas. Finalmente, se extrae una métrica de rendimiento para evaluar el modelo utilizando los datos de prueba y los datos predichos.

## Tecnologías Utilizadas

- **Python:** Lenguaje principal para scripting y análisis.
- **Jupyter Notebook:** Entorno interactivo para la realización del EDA y el modelado.
- **Base de Datos (PostgreSQL):** Para almacenar las predicciones y características.
- **Kafka:** Para la transmisión de datos.
- **Scikit-learn:** Biblioteca utilizada para el modelado y evaluación de modelos.

## Análisis Exploratorio de Datos (EDA) y Limpieza de Datos

Se comenzó con la carga y limpieza de los datos. Las columnas de los diferentes archivos CSV se estandarizaron para asegurar la consistencia entre los diferentes años.

### 2. Manejo de Valores Faltantes

Se llenaron los valores faltantes en las columnas numéricas con la media y en las columnas no numéricas con la moda.

### 3. Creación de Nuevas Características

- Interacción entre GDP y expectativa de vida saludable\*\*: Se creó una nueva característica multiplicando el PIB per cápita y la expectativa de vida saludable.
- **Transformaciones Logarítmicas:** Se aplicó el logaritmo natural a las características `GDP per capita` y `Social support` para reducir la variabilidad y normalizar los datos.

#### 4. Selección de Características

Basado en el análisis de correlación, se seleccionaron las siguientes características para el modelo:

- **log\_gdp\_per\_capita**
- **log\_social\_support**
- **healthy\_life\_expectancy**
- **freedom\_to\_make\_life\_choices**
- **generosity**
- **perceptions\_of\_corruption**
- **gdp\_health\_interaction**

Estas características se seleccionaron por su fuerte correlación con la variable de objetivo **happiness\_score**.

La variable **gdp\_health\_interaction** fue resultado de combinar las características **healthy\_life\_expectancy** y **gdp\_per\_capita** (las características que tenían más correlación según la matrix), esta característica mejoró los resultados del modelo en por lo menos un 5%

#### 5. Modelos de Predicción Evaluados

Se probaron varios modelos de **regresión** para encontrar el más adecuado:

- **Polynomial Features**
- **Random Forest**
- **Ridge Regression**
- **XGBoost**
- **Voting Regressor**
- **Stacking Regressor**

El modelo de **Stacking Regressor** fue el que mejor desempeño tuvo en términos de  $R^2$  y MSE con una asertividad aproximada del 79%.

#### 6. Flujo de Trabajo de Streaming de Datos

Se configuró un flujo de trabajo utilizando Kafka para transmitir los datos transformados. Los datos de prueba se enviaron al productor de Kafka, y el consumidor los recibió, aplicó el modelo entrenado para hacer predicciones y almacenó los resultados en una base de datos PostgreSQL.

## 7. Métrica de Rendimiento

Para evaluar el rendimiento del modelo, se utilizaron las métricas de  $R^2$  y el Error Cuadrático Medio (MSE). La evaluación se realizó tanto en los datos de prueba como en los datos predichos, garantizando así la robustez del modelo.

### Descripción de los Scripts utilizados

**Data Transformer:** Este script se encarga de cargar, limpiar y transformar los datos originales. Se estandarizaron las columnas, se llenaron los valores faltantes y se crearon nuevas características. Finalmente, se dividieron los datos en conjuntos de entrenamiento y prueba, y los datos de prueba se guardaron para su transmisión posterior con Kafka.

**Kafka Producer:** Este script envió los datos de prueba transformados a un tema de Kafka para su transmisión. Cada registro se envió como un mensaje JSON.

**Kafka Consumer:** Este script recibió los mensajes del productor de Kafka, aplicó el modelo entrenado para hacer predicciones y almacenó las predicciones junto con las características originales en una base de datos PostgreSQL. También se configuró para registrar los mensajes recibidos y las predicciones realizadas.

**Performance Metrics:** Este script cargó los datos de la base de datos y calculó las métricas de rendimiento (MSE y  $R^2$ ) para evaluar la precisión de las predicciones del modelo.

### Evidencias

#### resultado métricas de cada modelo probado:

##### Polynomial Features

- **Mean Squared Error (MSE):** 0.32471802472942013
- **$R^2$  Score:** 0.7484799653866189

##### Ridge Regression for Regularization

- **Mean Squared Error (MSE):** 0.362724635227172
- **$R^2$  Score:** 0.719040811228492

##### Random Forest

- **Mean Squared Error (MSE):** 0.2706934794858266
- **$R^2$  Score:** 0.7903262888266669

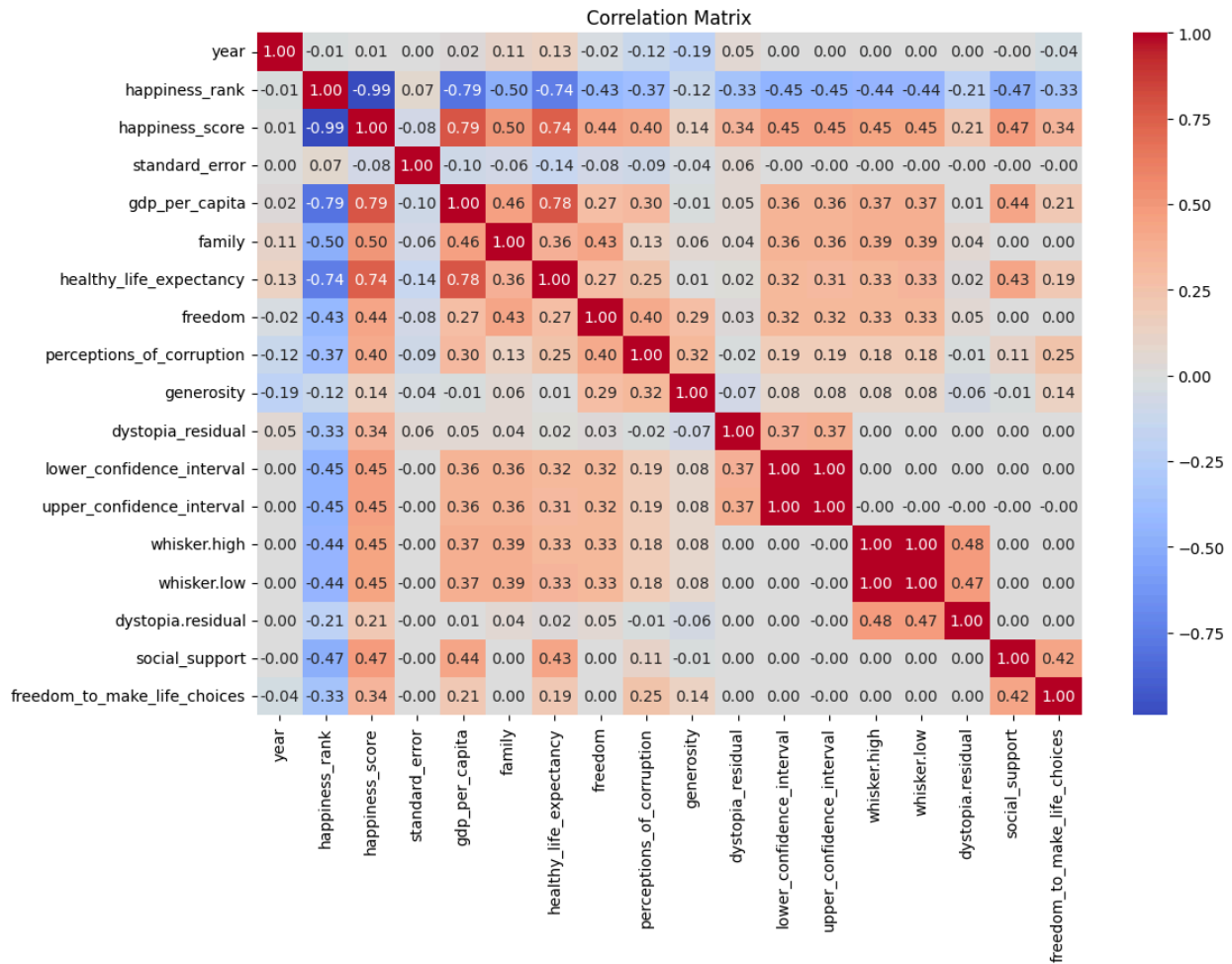
##### XGBoost

- **Mean Squared Error (MSE):** 0.2934685682195784
- **$R^2$  Score:** 0.7726851643112976

##### Stacking Regressor

- **Mean Squared Error (MSE):** 0.26137725175147697
- **$R^2$  Score:** 0.7975424509850881

El **mejor modelo** fue el Stacking Regressor con un **Mean Squared Error (MSE)** de 0.26137725175147697 y un  **$R^2$  Score** de 0.7975424509850881.



Matriz de correlación de los datos fuente

```
[21:47:03] isaac-opz ~/code/u/etl/ETL_Workshop3_Challenge
$ python3 scripts/performance_metric.py
/home/isaac-opz/.local/lib/python3.10/site-packages/scipy/
warnings.warn(f"A NumPy version >={np_minversion} and <{
Data loaded from database:
  year    country    region    happiness
0  2015    Serbia    Central and Eastern Europe
1  2015     Mali    Sub-Saharan Africa
2  2016    Rwanda    Sub-Saharan Africa
3  2018   Jamaica    Sub-Saharan Africa
4  2018 Tajikistan    Sub-Saharan Africa

[5 rows x 24 columns]
Mean Squared Error: 0.2407820680588461
R^2 Score: 0.8064558009665043
```

Métricas de performance del modelo de regresión (Datos sacados de la base de datos PSQL)

```
python3 scripts/kafka_consumer.py
2024-05-24 21:46:32.377 INFO - <BrokerConnection node_id=bootstrap-0 host=localhost:9092 <connected> [IPv4 ('127.0.0.1', 9092)]: Closing connection.
2024-05-24 21:46:32.477 INFO - (Re-)joining group happiness_group
2024-05-24 21:46:35.485 INFO - Elected group leader - performing partition assignments using range
2024-05-24 21:46:35.487 INFO - <BrokerConnection node_id=1 host=localhost:9092 <connecting> [IPv4 ('127.0.0.1', 9092)]: connecting to localhost:9092 [127.0.0.1, 9092] IPv4
2024-05-24 21:46:35.487 INFO - <BrokerConnection node_id=1 host=localhost:9092 <connecting> [IPv4 ('127.0.0.1', 9092)]: Connection complete.
2024-05-24 21:46:35.491 INFO - Successfully joined group happiness_group with generation 24
2024-05-24 21:46:35.491 INFO - Updated partition assignment: [TopicPartition(topic='happiness_topic', partition=0)]
2024-05-24 21:46:35.491 INFO - Setting newly assigned partitions [TopicPartition(topic='happiness_topic', partition=0)] for group happiness_group
2024-05-24 21:46:35.582 INFO - Received record: {'year': 2015, 'country': 'Serbia', 'region': 'Central and Eastern Europe', 'happiness_rank': 87, 'happiness_score': 5.123, 'standard_error': 0.04864, 'gdp_per_capita': 0.92053, 'family': 1.00964, 'healthy_life_expectancy': 0.74836, 'freedom': 0.20107, 'perceptions_of_corruption': 0.02617, 'generosity': 0.19231, 'dystopia_residual': 2.025, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.211025641, 'freedom_to_make_life_choices': 0.4235384615, 'gdp_health_interaction': 0.6888878368, 'log_gdp_per_capita': 0.6526011896, 'log_social_support': 0.7934564988}
2024-05-24 21:46:35.597 INFO - Features for prediction: log_gdp_per_capita log_social_support freedom_to_make_life_choices generosity perceptions_of_corruption gdp_health_interaction
0 0.652601 0.793456 0.423538 0.19231 0.02617 0.688888
2024-05-24 21:46:37.894 INFO - Inserted record into database: {'year': 2015, 'country': 'Serbia', 'region': 'Central and Eastern Europe', 'happiness_rank': 87, 'happiness_score': 5.123, 'standard_error': 0.04864, 'gdp_per_capita': 0.92053, 'family': 1.00964, 'healthy_life_expectancy': 0.74836, 'freedom': 0.20107, 'perceptions_of_corruption': 0.02617, 'generosity': 0.19231, 'dystopia_residual': 2.025, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.211025641, 'freedom_to_make_life_choices': 0.4235384615, 'gdp_health_interaction': 0.6888878368, 'log_gdp_per_capita': 0.6526011896, 'log_social_support': 0.7934564988, 'predicted_happiness_score': 5.139260287318993}
2024-05-24 21:46:37.894 INFO - Received record: {'year': 2015, 'country': 'Mali', 'region': 'Sub-Saharan Africa', 'happiness_rank': 138, 'happiness_score': 3.995, 'standard_error': 0.05602, 'gdp_per_capita': 0.26074, 'family': 1.03526, 'healthy_life_expectancy': 0.20583, 'freedom': 0.38857, 'perceptions_of_corruption': 0.12352, 'generosity': 0.18798, 'dystopia_residual': 1.79293, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.211025641, 'freedom_to_make_life_choices': 0.4235384615, 'gdp_health_interaction': 0.6536061142, 'log_gdp_per_capita': 0.2316988502, 'log_social_support': 0.7934564988}
2024-05-24 21:46:37.898 INFO - Features for prediction: log_gdp_per_capita log_social_support freedom_to_make_life_choices generosity perceptions_of_corruption gdp_health_interaction
0 0.231699 0.793456 0.423538 0.18798 0.12352 0.653606
2024-05-24 21:46:38.988 INFO - Inserted record into database: {'year': 2015, 'country': 'Mali', 'region': 'Sub-Saharan Africa', 'happiness_rank': 138, 'happiness_score': 3.995, 'standard_error': 0.05602, 'gdp_per_capita': 0.26074, 'family': 1.03526, 'healthy_life_expectancy': 0.20583, 'freedom': 0.38857, 'perceptions_of_corruption': 0.12352, 'generosity': 0.18798, 'dystopia_residual': 1.79293, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.211025641, 'freedom_to_make_life_choices': 0.4235384615, 'gdp_health_interaction': 0.6536061142, 'log_gdp_per_capita': 0.2316988502, 'log_social_support': 0.7934564988, 'predicted_happiness_score': 3.9907982151817976}
2024-05-24 21:46:38.988 INFO - Received record: {'year': 2016, 'country': 'Rwanda', 'region': 'Sub-Saharan Africa', 'happiness_rank': 152, 'happiness_score': 3.515, 'standard_error': 0.0478847468, 'gdp_per_capita': 0.32846, 'family': 0.61586, 'healthy_life_expectancy': 0.31805, 'freedom': 0.5432, 'perceptions_of_corruption': 0.05021, 'generosity': 0.23552, 'dystopia_residual': 0.96819, 'lower_confidence_interval': 3.444, 'upper_confidence_interval': 3.586, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.211025641, 'freedom_to_make_life_choices': 0.4235384615, 'gdp_health_interaction': 0.104663779, 'log_gdp_per_capita': 0.2842803766, 'log_social_support': 0.7934564988}
2024-05-24 21:46:38.988 INFO - Features for prediction: log_gdp_per_capita log_social_support freedom_to_make_life_choices generosity perceptions_of_corruption gdp_health_interaction
0 0.284281 0.793456 0.423538 0.23552 0.05021 0.104664
2024-05-24 21:46:39.028 INFO - Inserted record into database: {'year': 2016, 'country': 'Rwanda', 'region': 'Sub-Saharan Africa', 'happiness_rank': 152, 'happiness_score': 3.515, 'standard_error': 0.0478847468, 'gdp_per_capita': 0.32846, 'family': 0.61586, 'healthy_life_expectancy': 0.31805, 'freedom': 0.5432, 'perceptions_of_corruption': 0.05021, 'generosity': 0.23552, 'dystopia_residual': 0.96819, 'lower_confidence_interval': 3.444, 'upper_confidence_interval': 3.586, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.211025641, 'freedom_to_make_life_choices': 0.4235384615, 'gdp_health_interaction': 0.104663779, 'log_gdp_per_capita': 0.2842803766, 'log_social_support': 0.7934564988, 'predicted_happiness_score': 4.227562260820911}
2024-05-24 21:46:39.029 INFO - Received record: {'year': 2018, 'country': 'Jamaica', 'region': 'Sub-Saharan Africa', 'happiness_rank': 56, 'happiness_score': 5.89, 'standard_error': 0.0478847468, 'gdp_per_capita': 0.819, 'family': 0.990346641, 'healthy_life_expectancy': 0.603, 'freedom': 0.4028277145, 'perceptions_of_corruption': 0.031, 'generosity': 0.096, 'dystopia_residual': 2.212031619, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.493, 'freedom_to_make_life_choices': 0.575, 'gdp_health_interaction': 0.507567, 'log_gdp_per_capita': 0.909346641, 'log_social_support': 0.9134808045, 'predicted_happiness_score': 5.22760957468489}
2024-05-24 21:46:39.032 INFO - Features for prediction: log_gdp_per_capita log_social_support freedom_to_make_life_choices generosity perceptions_of_corruption gdp_health_interaction
0 0.909347 0.913481 0.575 0.096 0.031 0.507567
2024-05-24 21:46:39.436 INFO - Inserted record into database: {'year': 2018, 'country': 'Jamaica', 'region': 'Sub-Saharan Africa', 'happiness_rank': 56, 'happiness_score': 5.89, 'standard_error': 0.0478847468, 'gdp_per_capita': 0.819, 'family': 0.990346641, 'healthy_life_expectancy': 0.603, 'freedom': 0.4028277145, 'perceptions_of_corruption': 0.031, 'generosity': 0.096, 'dystopia_residual': 2.212031619, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.493, 'freedom_to_make_life_choices': 0.575, 'gdp_health_interaction': 0.507567, 'log_gdp_per_capita': 0.909346641, 'log_social_support': 0.9134808045, 'predicted_happiness_score': 5.22760957468489}
2024-05-24 21:46:39.437 INFO - Received record: {'year': 2018, 'country': 'Tajikistan', 'region': 'Sub-Saharan Africa', 'happiness_rank': 88, 'happiness_score': 5.199, 'standard_error': 0.0478847468, 'gdp_per_capita': 0.474, 'family': 0.990346641, 'healthy_life_expectancy': 0.598, 'freedom': 0.4028277145, 'perceptions_of_corruption': 0.034, 'generosity': 0.1166, 'dystopia_residual': 2.212031619, 'lower_confidence_interval': 5.2823949045, 'upper_confidence_interval': 5.4819745223, 'whisker_high': 5.4523257175, 'whisker_low': 5.2537129941, 'dystopia_residual': 1.8502378056, 'social_support': 1.166, 'freedom_to_make_life_choices': 0.292, 'gdp_health_interaction': 0.283452, 'log_gdp_per_capita': 0.387979793, 'log_social_support': 0.7728821460}
2024-05-24 21:46:39.443 INFO - Features for prediction: log_gdp_per_capita log_social_support freedom_to_make_life_choices generosity perceptions_of_corruption gdp_health_interaction
0 0.387979 0.772882 0.292 0.167 0.034 0.283452
```

kafka consumer recibiendo datos y enviándolos a la base de datos

Tables										
50 rows • 694ms										
database: etl_workshop3	year	country	region	happiness_rank	happiness_score	standard_error	gdp_per_capita	family	he	
schema: public	2015	Serbia	Central and	87	5.123	0.04864	0.92053	1.00964	0.	
Search tables	2015	Mali	Sub-Saharan	138	3.995	0.05602	0.26074	1.03526	0.	
	2016	Rwanda	Sub-Saharan	152	3.515	0.0478847468	0.32846	0.61586	0.	
happiness_predictions	2018	Jamaica	Sub-Saharan	56	5.89	0.0478847468	0.819	0.990346641	0.	
	2018	Tajikistan	Sub-Saharan	88	5.199	0.0478847468	0.474	0.990346641	0.	
	2016	Croatia	Central and	74	5.488	0.0478847468	1.18649	0.60809	0.	
	2018	Uzbekistan	Sub-Saharan	44	6.096	0.0478847468	0.719	0.990346641	0.	
	2017	United Kingd...	Sub-Saharan	19	6.7140082251	0.0478847468	1.4416339397	1.4964080081	0.	
	2019	France	Sub-Saharan	24	6.092	0.0478847468	1.324	0.990346641	1.	
	2018	Cameroon	Sub-Saharan	99	4.975	0.0478847468	0.535	0.990346641	0.	
	2019	Tanzania	Sub-Saharan	153	3.231	0.0478847468	0.476	0.990346641	0.	
	2015	Haiti	Latin Americ...	119	4.518	0.07331	0.26673	0.74302	0.	
	2019	Taiwan	Sub-Saharan	25	6.446	0.0478847468	1.368	0.990346641	0.	
	2016	South Korea	Eastern Asia	57	5.835	0.0478847468	1.35948	0.72194	0.	
	2019	Rwanda	Sub-Saharan	152	3.334	0.0478847468	0.359	0.990346641	0.	
	2017	Saudi Arabia	Sub-Saharan	37	6.3439998627	0.0478847468	1.5306235552	1.286677599	0.	
	2019	Brazil	Sub-Saharan	32	6.3	0.0478847468	1.004	0.990346641	0.	
	2015	Nigeria	Sub-Saharan	78	5.268	0.04192	0.65435	0.90432	0.	
	2019	Northern Cyp...	Sub-Saharan	64	5.718	0.0478847468	1.263	0.990346641	1.	
	2017	Costa Rica	Sub-Saharan	12	7.0789999962	0.0478847468	1.1097062826	1.4164036512	0.	
	2019	Estonia	Sub-Saharan	55	5.893	0.0478847468	1.237	0.990346641	0.	
	2015	Italy	Western Euro...	58	5.948	0.03914	1.25114	1.19777	0.	
	2016	Bulgaria	Central and	129	4.217	0.0478847468	1.11306	0.92542	0.	

base de datos con registros enviados por Kafka Consumer