Final Report

Mariana Cifuentes Rivera

Universidad Autónoma de Occidente

Data Engineering and Artificial Intelligence

Subject: ETL

Teacher: Breyner Posso Bautista

November 2025

Final Report
ETL Process using Apache Kafka and Machine Learning

1. Dataset Description

The dataset used in this project corresponds to the *World Happiness Report*, which annually collects information about happiness levels in different countries around the world.
For this workshop, data from **2015 to 2019** was used, including economic, social, and well-being indicators.
Each file contains variables that explain the factors that determine a country's *Happiness Score*, measured on a scale from 0 to 10.

Main variables:

- **Country:** Name of the country.
- **Year:** Year of record (2015–2019).
- **GDP per Capita:** Contribution of per-capita GDP to the happiness score.
- **Social support:** Level of perceived social support.
- **Healthy life expectancy:** Contribution of healthy life expectancy to happiness.
- **Freedom:** Perceived freedom to make life decisions.
- **Generosity:** Level of generosity in the population.
- **Perceptions of corruption:** Level of perceived corruption in institutions.
- **Score (Happiness Score):** Target variable to predict.

Data cleaning and standardization:

- Column names were standardized across years.
- The five datasets were concatenated into a single DataFrame.
- Redundant information such as *Rank*, *Dystopia Residual*, and other indicators derived from *Score* was removed to avoid data leakage.
- A linear interpolation was applied to the *Perceptions of corruption* column to replace a single missing value.
- Zero values in variables such as *GDP per Capita* and *Healthy life expectancy* were confirmed as valid and meaningful, not missing data.

2. Main Findings from the EDA

The exploratory data analysis (EDA) helped to understand the structure and behavior of the variables:

- **GDP per Capita**, **Healthy life expectancy**, and **Social support** showed the highest correlations with the *Happiness Score* ($r > 0.7$).
- **Freedom** and **Perceptions of corruption** had moderate correlations ($r \approx 0.5$ and 0.4, respectively).
- **Generosity** had a weaker but still positive correlation ($r \approx 0.15$), suggesting some influence on well-being perception.
- No invalid values or anomalous outliers were detected; zeros represent valid cases of null contribution to the score.
- The *Score* variable followed an almost normal distribution centered around 5.3.

EDA conclusion:

The dataset has high quality and internal consistency. The explanatory variables are independent, conceptually meaningful, and statistically significant for building a linear regression model.

3. Model Selection and Training Process

After cleaning, the final dataset included the following predictors:

- GDP per Capita
- Social support
- Healthy life expectancy
- Freedom
- Generosity
- Perceptions of corruption

Data split:

- 70% training set
- 30% test set

3.1 Evaluated Models and Final Model Selection

Four regression algorithms from the linear family were tested. These models were chosen because the global happiness data shows mostly proportional and additive relationships among variables (for example, higher GDP or social support tends to increase the happiness

score). Therefore, a linear approach was the most appropriate without introducing unnecessary complexity.

Model	Description	Purpose
Multiple Linear Regression (OLS)	Fits a linear equation between predictors and the happiness score.	Baseline model; easy to interpret each factor's influence.
Ridge Regression (L2)	Similar to linear regression but adds a small penalty to reduce multicollinearity.	Tested in case of correlations between variables like GDP and life expectancy.
Lasso Regression (L1)	Applies a penalty that forces less relevant coefficients to zero.	Tested to see if some variables could be removed without affecting performance.
Elastic Net (L1 + L2)	Combines both penalties for balance.	Used to find a midpoint between Ridge (stability) and Lasso (simplicity).

Performance results (test set):

Model	R^2	MAE	RMSE
Linear Regression	**0.726480457058148**	**0.448639645400341**	**0.584383724574775**
Ridge (α = 0.001)	0.726480388493964	0.448639712624447	0.584383797819616
Elastic Net (α = 0.001, l1 = 0.1)	0.726439367418602	0.448682950614655	0.584427617626438
Lasso (α = 0.001)	0.726391456621898	0.448753608405005	0.584478793058074

Analysis and interpretation:

All four models performed almost identically, but the **Multiple Linear Regression** achieved the best results in every metric, even if only by small decimals.

- It reached the highest **R^2 (0.726480)**, meaning it explains **72.6%** of the variability in happiness scores.
- It obtained the lowest **MAE (0.4486396)** and **RMSE (0.5843837)**, showing the most accurate predictions overall.
- Ridge, Lasso, and Elastic Net produced nearly the same results, but always slightly worse.

During cross-validation, the optimal regularization parameter **α for Ridge was 0.001**, which is **very close to zero**, meaning the penalty had almost no effect.

This indicates that Ridge behaved almost exactly like an OLS regression, confirming there was **no serious multicollinearity** and that regularization was unnecessary.

Additional validation showed:

- **VIF values (1.18–2.63)** indicated low correlations among predictors.
- Residual plots and statistical tests confirmed compliance with classical linear regression assumptions (linearity, homoscedasticity, normality, independence).
- No overfitting was found: **R^2_train = 0.778** and **R^2_test = 0.726**, a small difference of about **5%**.

Conclusion:

The **Multiple Linear Regression model** was selected as the final model because:

- It had the best quantitative performance across all metrics.
- The Ridge α value was almost null, so the penalty added no improvement.
- It maintains fully interpretable coefficients, clearly showing each variable's effect on happiness.
- It satisfies all statistical assumptions and generalizes well to unseen data.

4. Model Evaluation and Visualization of Performance

An interactive dashboard built with **Plotly Dash** displays the main performance metrics (KPIs) and visual diagnostics of the model.

4.1 Key KPIs (Test set)

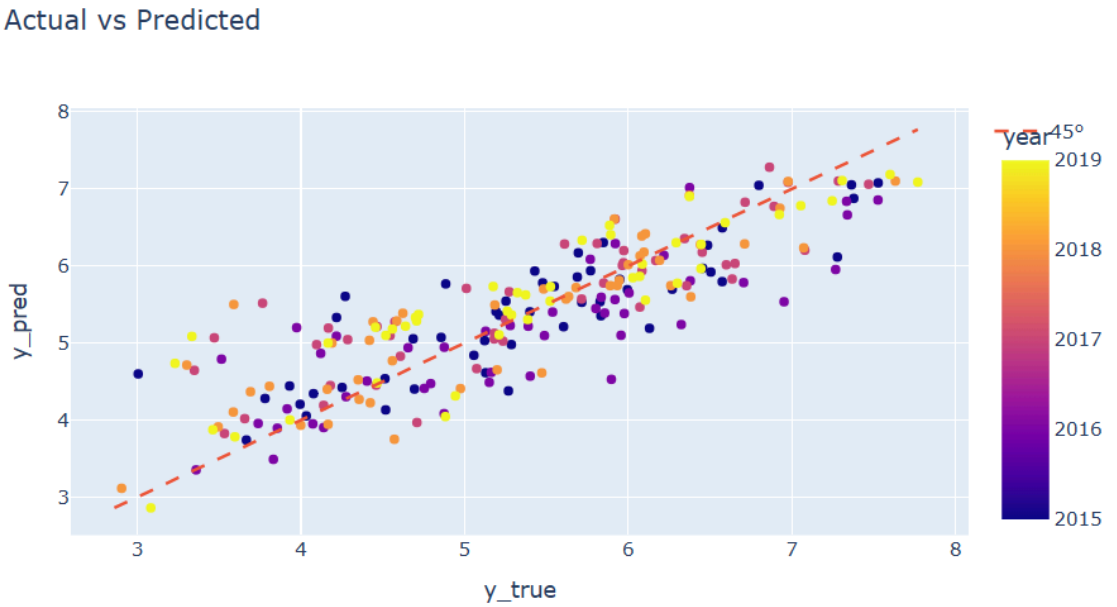
Metric	Value	Interpretation
R^2	0.726	The model explains 72.6% of the variance.
MAE	0.449	Average error of ±0.45 points.
RMSE	0.584	Moderate average squared error.
MAPE	9.16%	Low percentage error (<10%).
Bias (ŷ−y)	+0.038	Slight overestimation.
Durbin–Watson	2.149	No residual autocorrelation.
# Test Records	235	Adequate sample size.

General conclusion:

The model demonstrates strong explanatory power, small errors, and consistent statistical assumptions, making it reliable and stable.

4.2 Graphs and Interpretation

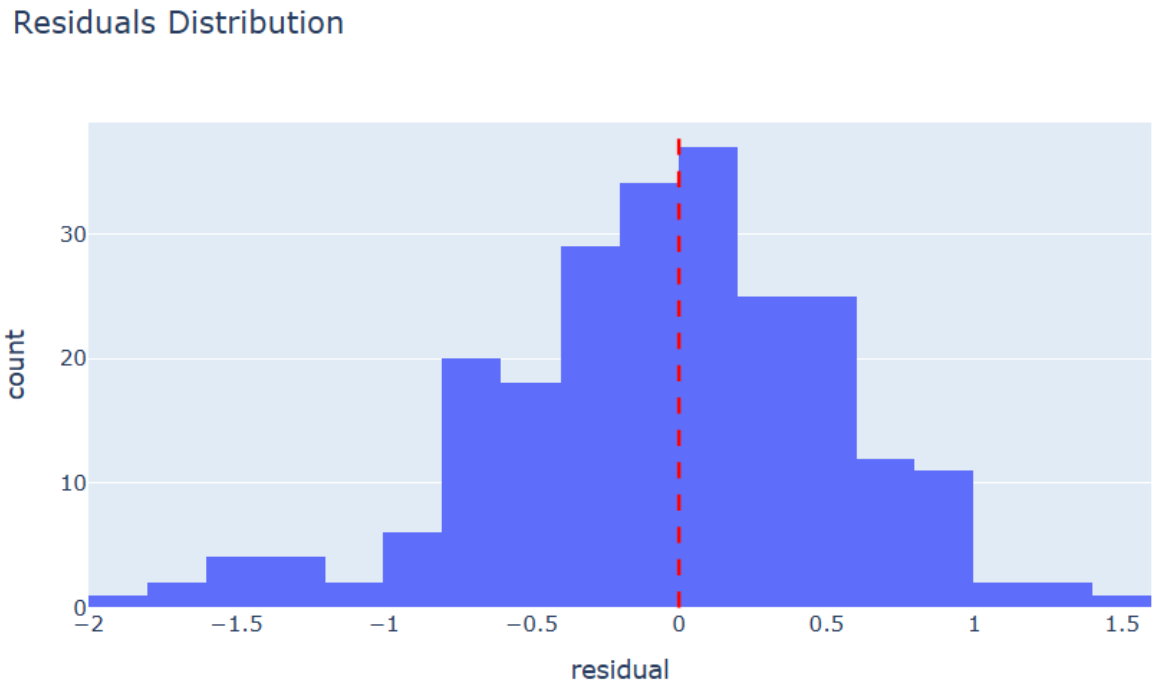
a) Actual vs Predicted



Points align closely to the 45° line, showing good agreement between actual and predicted values.

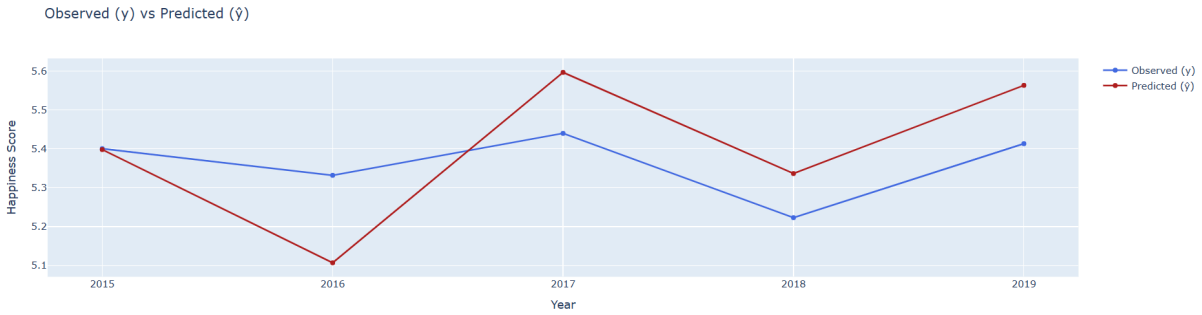
There are no temporal biases, confirming stable behavior between 2015 and 2019.

b) Residual Distribution



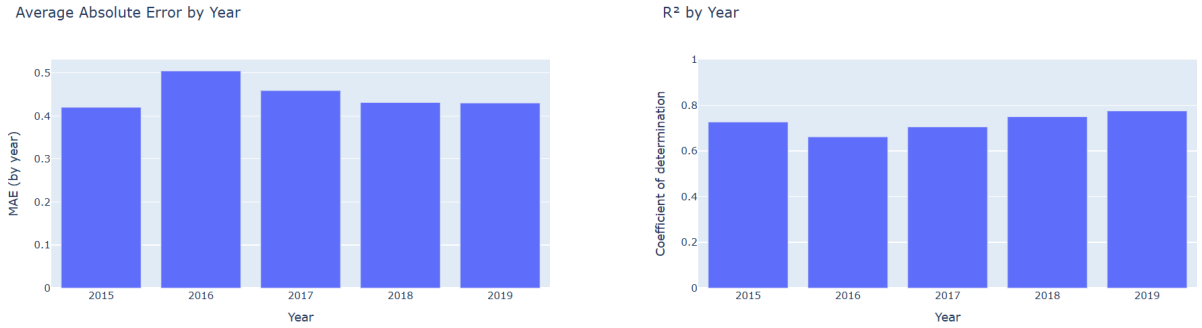
The residual histogram is approximately symmetric and centered at zero, showing random errors consistent with normality and independence.

c) Observed vs Predicted Comparison (by Year)



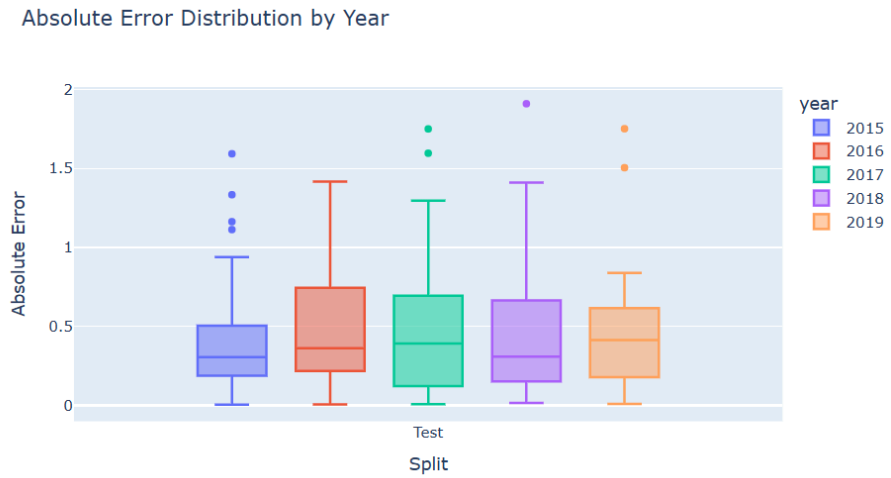
The model accurately follows the overall global happiness trend.
It only slightly underestimates the 2016 value (≈0.22 points) but remains consistent across all other years.

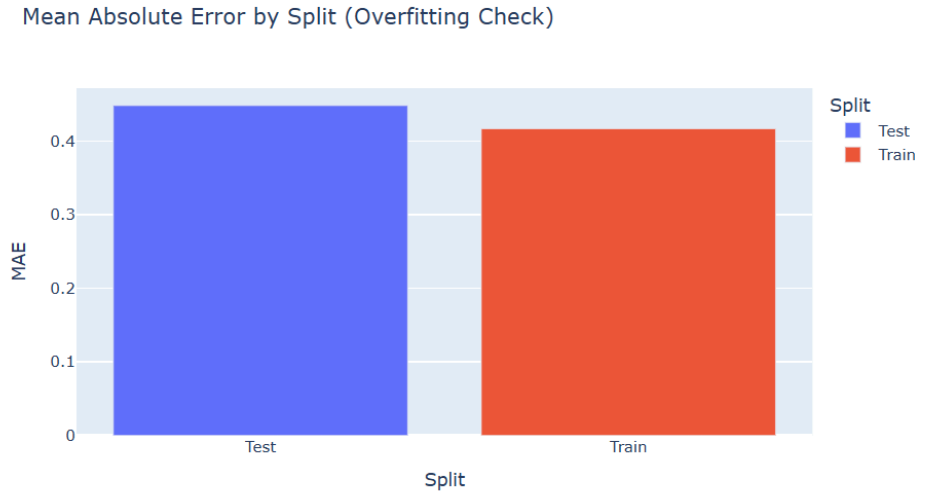
d) Mean Absolute Error and R² by Year



The annual MAE ranges between 0.42–0.50, and R² remains between 0.67–0.79, showing temporal consistency and model robustness.

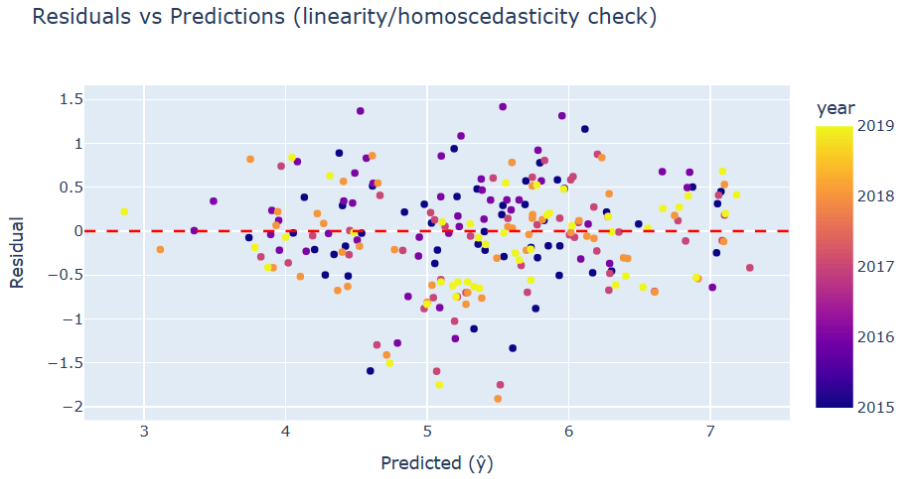
e) Boxplot of Errors + Overfitting Check

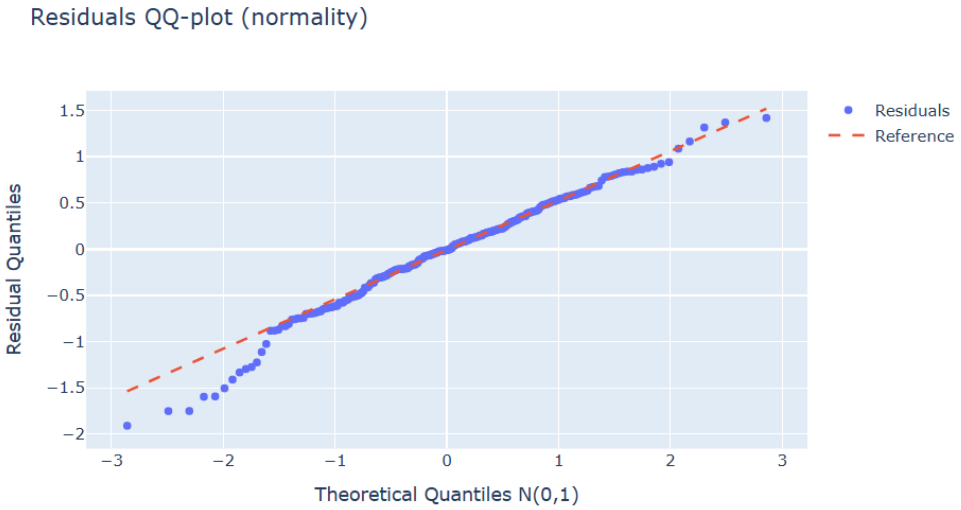




Error dispersion is stable across years, and training and test errors are almost identical, confirming no overfitting.

f) Residuals vs Predictions + QQ-Plot

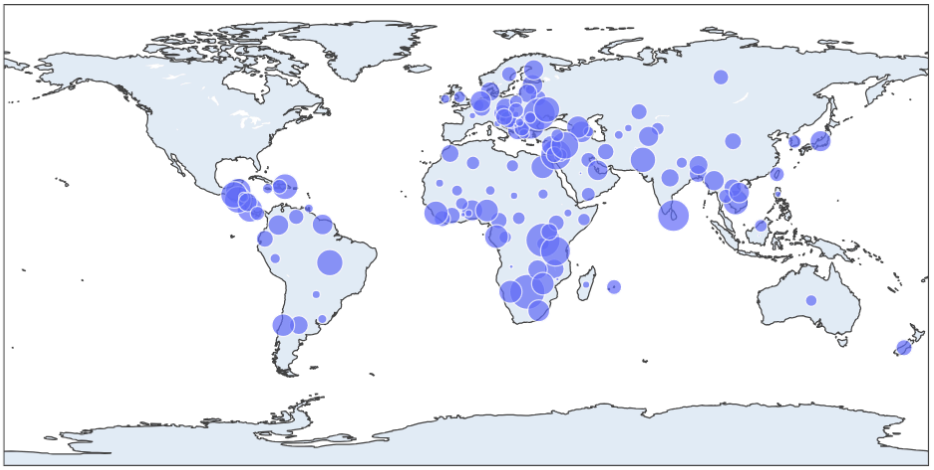




Residuals are randomly distributed (linearity and homoscedasticity) and approximately follow a normal distribution, confirming the model's assumptions.

g) MAE Map by Country + Top-10 Errors + VIF

Mean Absolute Error by Country (MAE)



Top-10 Absolute Errors

Country	Year	Actual (y)	Predicted (ŷ)	|Error|	Residual
botswana	2018	3.5900	5.4998	1.9098	-1.9098
rwanda	2019	3.3340	5.0855	1.7515	-1.7515
botswana	2017	3.7660	5.5164	1.7504	-1.7504
rwanda	2017	3.4710	5.0673	1.5963	-1.5963
syria	2015	3.0060	4.5988	1.5928	-1.5928
tanzania	2019	3.2310	4.7357	1.5047	-1.5047
brazil	2016	6.9520	5.5347	1.4173	1.4173
tanzania	2018	3.3030	4.7139	1.4109	-1.4109
moldova	2016	5.8970	4.5284	1.3686	1.3686
sri lanka	2015	4.2710	5.6046	1.3336	-1.3336

The model fits better in the Americas and Europe but shows larger errors in Africa (Botswana, Rwanda, Tanzania, etc.), where economic factors alone do not fully explain subjective well-being.

VIF (Multicollinearity)

Variable	VIF
gdp_per_capita	2.629
healthy_life_expectancy	2.496
social_support	1.792
freedom	1.521
perceptions_of_corruption	1.380
generosity	1.181

All **VIF values < 3**, confirming the absence of multicollinearity.

5. Discussion of the Streaming Process (Kafka)

Kafka Producer
The producer script (*producer_happiness.py*):

- Reads and transforms the CSV files (2015–2019) applying the same EDA transformations.
- Standardizes columns, adds the year, and normalizes country names.
- Sends each record serialized as JSON to the Kafka topic **happiness-topic**, simulating real-time streaming.

Kafka Consumer
The consumer script (*consumer_happiness.py*):

- Listens to the **happiness-topic**.
- Loads the trained model (.pkl) and generates predictions in real time.
- Inserts data into a **MySQL star schema** with the following tables:
    - **dim_country**
    - **dim_time**
    - **fact_predictions**

This setup allows historical predictions to be stored and later analyzed using dashboards or BI tools.

6. Conclusions

- The **Multiple Linear Regression model** explained **72.6%** of the variability in happiness scores for unseen data, with low average errors (MAE = 0.45, RMSE = 0.58).

- The classical regression assumptions were verified: **linearity, homoscedasticity, independence, and normality of residuals**.

- Geographic analysis showed strong predictions in most regions and localized errors in countries affected by non-socioeconomic factors.

- No overfitting or multicollinearity issues were detected (**VIF < 3**).

- The streaming system implemented with **Apache Kafka** successfully simulated real-time data flow and stored model predictions in a relational database.

Overall, the project integrates all components of a complete data workflow (**ETL, Machine Learning, and Data Streaming**), demonstrating how predictive analytics can be deployed in near-real-time environments to analyze global well-being data.