Workshop 003

Happiness Score Prediction and Data Streaming

Realizado por:

Luis Angel Garcia(2230177)

DOCENTE: JAVIER ALEJANDRO VERGARA ZORRILLA

ETL

PROGRAMA DE INGENIERÍA DE DATOS E INTELIGENCIA ARTIFICIAL

FACULTAD DE INGENIERÍA

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE

SANTIAGO DE CALI

2025

1. Introduction

This report presents the development of a comprehensive machine learning and real-time data streaming system designed to predict the Happiness Score of various countries based on socioeconomic indicators. The project was carried out as part of Workshop 3 and integrates key components such as exploratory data analysis (EDA), supervised regression modeling, metric evaluation, and the deployment of a continuous data flow architecture using modern technologies including Apache Kafka and PostgreSQL.

The foundation of the system lies in the annual World Happiness Reports published by the United Nations between 2015 and 2019. These datasets include critical indicators such as Gross Domestic Product (GDP) per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. An ETL (Extract, Transform, Load) pipeline was implemented to unify, clean, and standardize the data from different years in preparation for analysis and modeling.

Subsequently, several regression models were trained and evaluated using metrics such as the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The final stage involved deploying a real-time prediction system using Apache Kafka to simulate data streaming and PostgreSQL to store incoming predictions in a structured format.

This document outlines each phase of the project in detail, highlighting key technical decisions, evaluation outcomes, and areas for future improvement.

2. Technologies Implemented

The implementation of this project required the integration of various tools and technologies, each selected based on its suitability for data processing, modeling, real-time streaming, and reproducibility. The key components are outlined below:

2.1 Programming Language and Data Science Libraries

- **Python 3.12** was used as the primary programming language due to its extensive ecosystem and community support in the field of data science.
- **Pandas** facilitated data cleaning, transformation, and aggregation across all five datasets.
- **Scikit-learn** was employed for feature scaling, model training, and evaluation. It also provided tools such as Ridge, HuberRegressor, and performance metrics like RMSE and R^2.
- **Joblib** was used to serialize the trained model and scaler for inference purposes.

2.2 Data Streaming

- **Apache Kafka** served as the core streaming platform. A KafkaProducer was used to send preprocessed data into a topic (happiness_topic), while a KafkaConsumer received and processed the messages in real time.
- **The kafka-python libra**ry provided Python bindings to interact with Kafka brokers seamlessly.

2.3 Database Management

- **PostgreSQL** was chosen as the target database for storing predictions and feature values. It offers reliability, structured querying (SQL), and integration with Python via psycopg2.
- A schema was created automatically by the consumer to store fields such as GDP, health, freedom, and predicted happiness scores.

2.4 Containerization and Deployment

- Docker and Docker Compose were utilized to encapsulate and orchestrate the application. Services included:
  - A Kafka broker and Zookeeper
  - A PostgreSQL container with volume mapping
  - An application container (app) responsible for running the consumer logic
  - This setup enabled environment isolation, quick reproducibility, and simplified deployment without the need for manual installation of services.

2.5 Development Environment

- The project was primarily developed and tested on Ubuntu WSL under a virtual Python environment (venv), ensuring compatibility and avoiding dependency conflicts.
- **Jupyter Notebooks** were used during the EDA and model training phases to allow for interactive exploration and visualization.

3. Dataset Description

The dataset used in this project is derived from the World Happiness Reports published annually by the United Nations Sustainable Development Solutions Network. These reports contain cross-country measurements of happiness based on survey responses and aggregated indicators that reflect both economic and social well-being. For this project, data from five consecutive years—2015 to 2019—were used, each provided as an independent CSV file. Each dataset includes a range of variables such as:

- Country name
- Happiness Score (or Score): the target variable, measuring self-reported well-being on a scale from 0 to 10
- GDP per capita: economic prosperity
- Social support: presence of supportive relationships
- Healthy life expectancy
- Freedom to make life choices
- Generosity
- Perceptions of corruption (Trust)
- Dystopia Residual: a benchmark representing hypothetical worst-case scenarios.

However, the datasets presented structural inconsistencies. Variable names changed from year to year (e.g., *"Economy (GDP per Capita)"* in 2015 versus *"GDP per capita"* in 2018), some

variables were renamed or removed, and different years had different numbers of features and countries. For instance, the 2016 dataset included confidence intervals for each score, while these were absent in others. Additionally, some datasets used alternative naming conventions such as dots instead of spaces.
To enable proper integration and analysis, these inconsistencies were addressed through a structured preprocessing phase that involved standardizing column names, aligning features across years, and appending a Year column to retain temporal context.

4. ETL Process

A complete ETL (Extract, Transform, Load) pipeline was implemented in Python and orchestrated using Apache Airflow, which enabled structured and automated execution of each step. The goal of this pipeline was to resolve inconsistencies across yearly datasets, handle missing data intelligently, and produce a standardized dataset suitable for modeling.

Extraction

Five CSV files, each corresponding to one year between 2015 and 2019, were ingested using Python's pandas library. During this phase, exploratory checks were performed to identify inconsistencies in feature names, data types, and variable presence. For example:

- "Economy (GDP per Capita)" in 2015 vs. "GDP per capita" in 2018–2019
- "Health (Life Expectancy)" vs. "Healthy life expectancy"
- "Trust (Government Corruption)" vs. "Perceptions of corruption"

Some years included additional metadata like confidence intervals, while others omitted such variables altogether. To unify the data across years, a preprocessing logic was required.

Transformation

The transformation stage involved the following actions:
Column normalization: All columns were renamed to consistent labels such as GDP, Social support, Health, Freedom, Trust, Generosity, Dystopia, and Year.
Column filtering: Redundant variables like Region or error margins were discarded.
Missing value treatment:
The variable Trust was imputed using the global mean.
Dystopia values were imputed based on the mean per year to preserve internal consistency.
Time annotation: A Year column was added to each dataset to capture its temporal context.
Concatenation: The cleaned and transformed datasets from all five years were merged into one unified dataset.
This entire logic was encapsulated into a DAG (Directed Acyclic Graph) in Apache Airflow, enabling modular, traceable, and repeatable execution of the ETL pipeline.

Load

The final dataset, named happiness_cleaned.csv, contains 782 rows and 11 consistently named columns. It was saved locally in the /data directory and used as input for model training, streaming, and evaluation phases.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out to understand the underlying patterns, distributions, and correlations among the variables in the unified dataset. This step was

essential to assess the quality of the data, identify relevant features for modeling, and detect any irregularities or biases that could influence predictive **performance.**

5.1 Descriptive Statistics and Distributions

Descriptive statistics were calculated for all numerical features, revealing that most variables followed reasonably normal distributions, with a few exceptions:
GDP showed a positively skewed distribution, indicating that a few countries have significantly higher income levels.
Generosity and Trust displayed high variance and outliers, reflecting diverse perceptions and cultural behaviors across nations.
The target variable, Happiness Score, had a bell-shaped distribution concentrated between scores of 4 and 7.
To visualize these findings, histograms were plotted for each feature. These plots confirmed that some indicators such as Freedom and Health were clustered tightly, whereas variables like Trust had widespread variability**.**

5.2 Yearly Distribution Overview

The dataset spans five years, from 2015 to 2019. By plotting the distribution of the Happiness Score across years, a consistent global pattern was observed: the average score remained relatively stable, with Scandinavian countries consistently ranking among the highest. The number of countries represented in each year was similar, ensuring temporal balance in the dataset.

5.3 Correlation Analysis

A Pearson correlation matrix was generated to assess linear relationships between variables. The results, visualized using a heatmap, showed that:
Social support, GDP, and Health had the strongest positive correlations with Happiness Score.
Trust and Generosity had lower correlation values but were retained due to their conceptual relevance.
The variable Dystopia also showed moderate positive correlation, representing the residual component of well-being not explained by other metrics.
These insights guided the feature selection process and validated the inclusion of all eight predictors in the modeling phase.

5.4 Visual Summaries

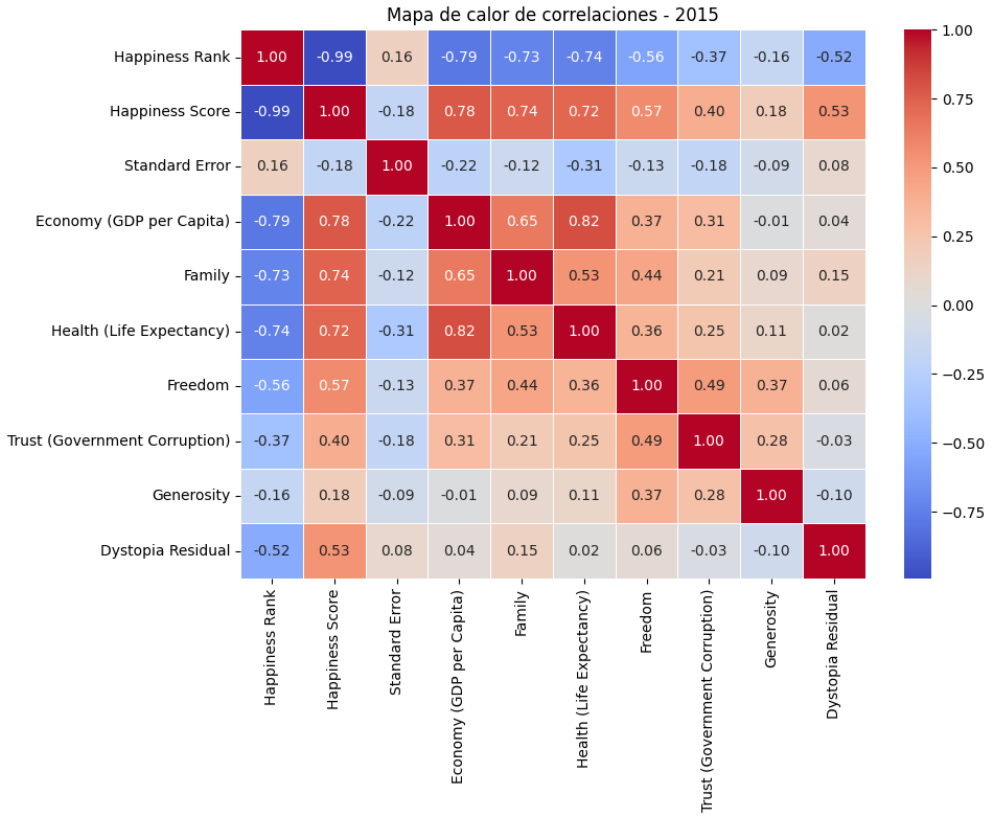
Two key visualizations were produced as part of the EDA:
A heatmap of the correlation matrix, highlighting the most influential features.
Distribution plots for each numeric variable, categorized by year and overall.
These visual summaries confirmed the dataset's suitability for supervised regression modeling and highlighted areas where outlier handling or feature scaling would be necessary.

Correlation Analysis of Variables (2015–2019)

To understand the relationships between the features and their predictive potential for the Happiness Score, Pearson correlation matrices were generated for each year from 2015 to 2019. These heatmaps visually depict the linear correlation between variables, where values closer to 1 or -1 indicate strong positive or negative relationships, respectively.

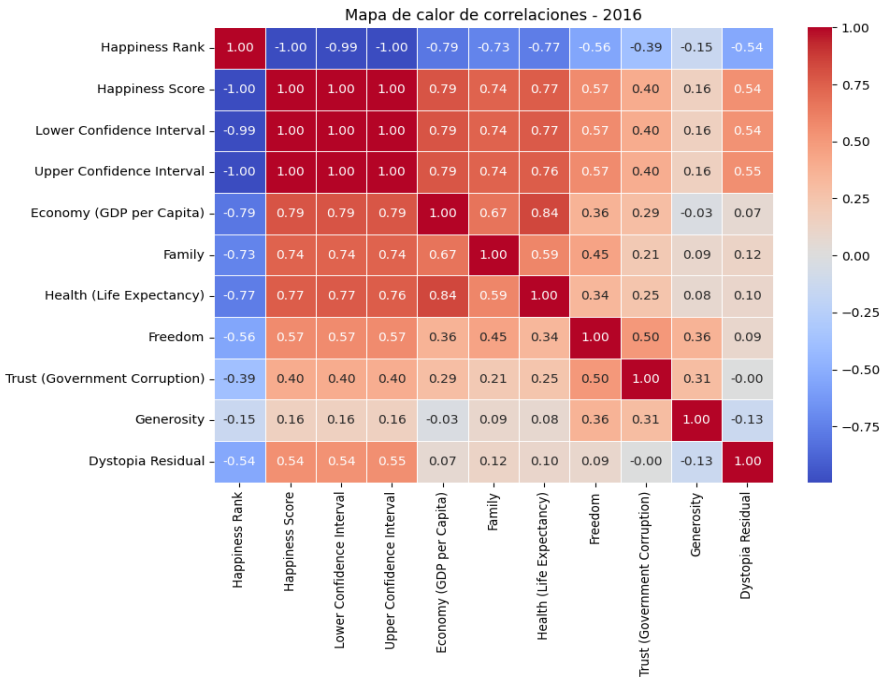
2015

The Happiness Score showed a strong positive correlation with GDP per capita (0.78), Health (0.72), and Family (0.74). This confirms that economic development, life expectancy, and social support were major contributors to happiness in this period. Negative correlations were observed with Happiness Rank (-0.99), as expected since a lower rank indicates higher happiness.



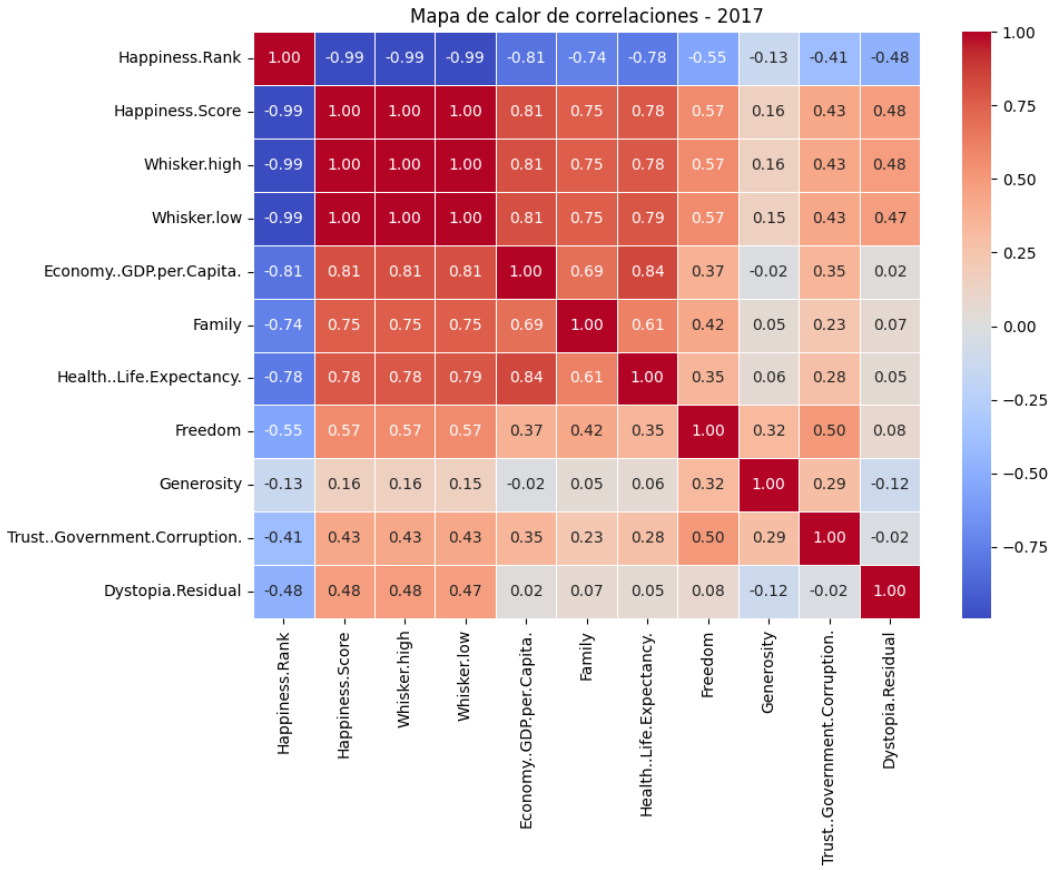
2016

The correlations remained consistent, with GDP per capita (0.79), Health (0.77), and Family (0.74) again showing strong alignment with the Happiness Score. Additionally, new confidence interval variables (upper/lower) were introduced, displaying near-perfect correlation with the score, suggesting they were derived metrics and thus excluded during feature selection.



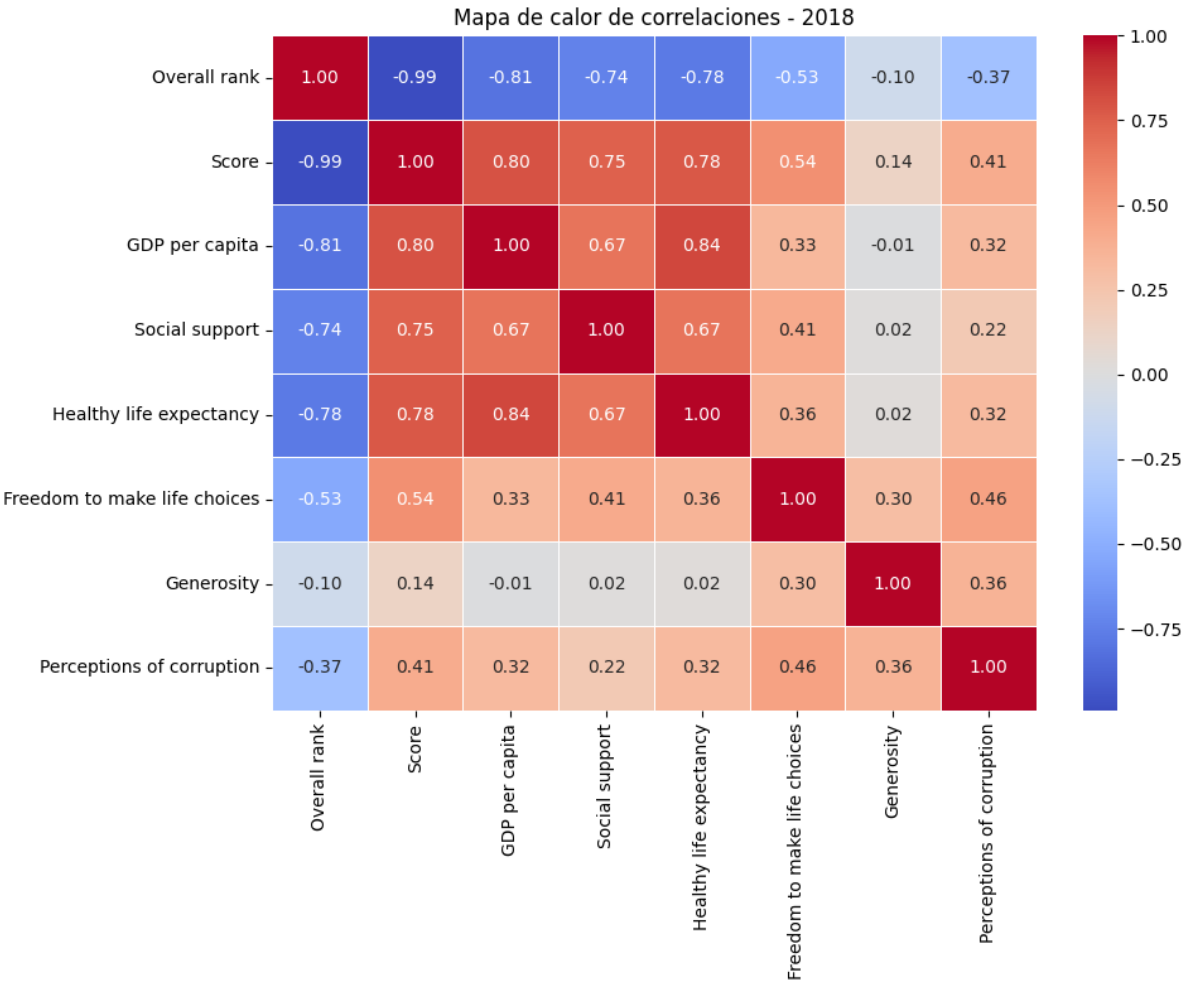
2017

Similar trends were observed with GDP (0.81), Health (0.78), and Family (0.75) showing strong associations. Variables such as Freedom (0.57) and Trust in government (0.43) began to gain prominence, showing that beyond economic factors, perceived institutional trust and autonomy contributed to well-being.



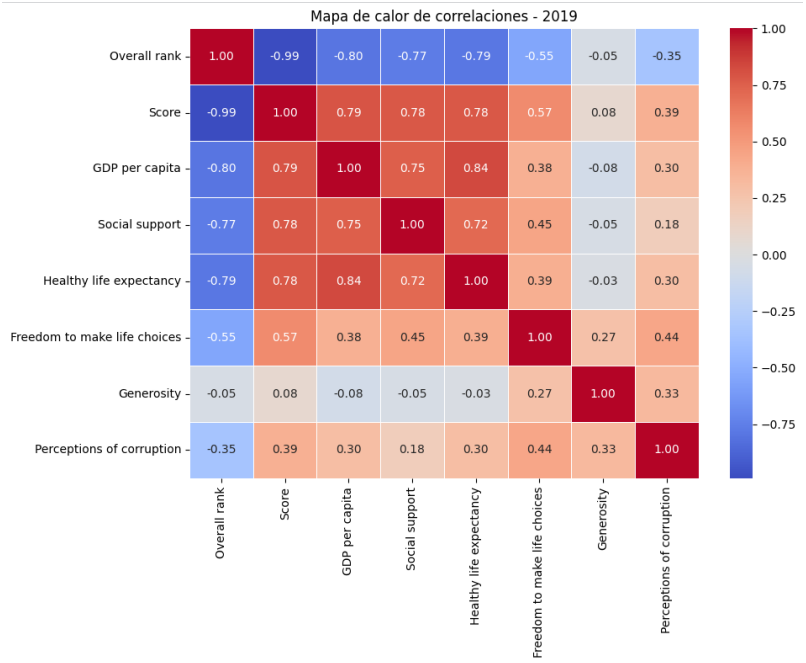
2018

While variable names were slightly modified (e.g., “Freedom to make life choices”), the relationships remained stable. GDP per capita (0.80) and Healthy life expectancy (0.78) maintained high correlation with the Happiness Score. Notably, Freedom (0.54) and Perceptions of Corruption (0.41) saw increases in influence, indicating societal factors became more important.



2019

The 2019 matrix further reinforced these insights, with GDP per capita (0.79), Social support (0.78), and Healthy life expectancy (0.78) staying as top contributors. Freedom (0.57) and Perceptions of Corruption (0.39) retained relevance, emphasizing how personal and institutional freedoms increasingly mattered to global happiness perceptions.



Distribution Analysis of Scaled Numerical Variables (2015–2019)
The boxplots generated for each year (2015 to 2019) present a scaled view of the distribution of numerical variables, allowing for a standardized comparison across features with different units and scales. This approach is particularly useful for visualizing skewness, variability, and outlier behavior across years.

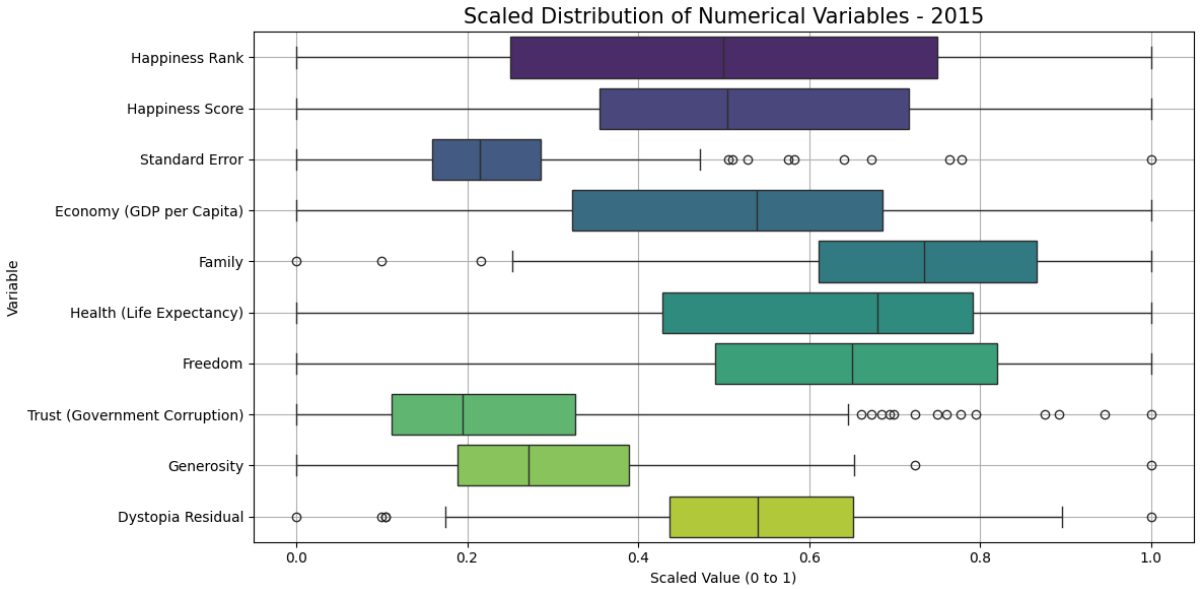
Across all years, variables such as Health (Life Expectancy) and Social Support exhibited high median values with relatively tight interquartile ranges, suggesting consistency and strong contributions to the overall happiness score. Meanwhile, Trust in Government Corruption and Generosity tended to show lower medians and greater variability, indicating greater dispersion and potential contextual influence depending on the country and time.

Outliers were particularly notable in Generosity and Trust, consistent with heterogeneous socio-political conditions across nations. In contrast, GDP per Capita consistently displayed a wider distribution but remained fairly symmetric year over year, reflecting global economic disparities.

The use of scaled values from 0 to 1 ensured comparability, enabling insights into which variables were systematically more or less influential across all datasets. Overall, this standardized distribution analysis supported robust feature selection for the modeling stage and helped in identifying consistent trends for high-impact predictors.

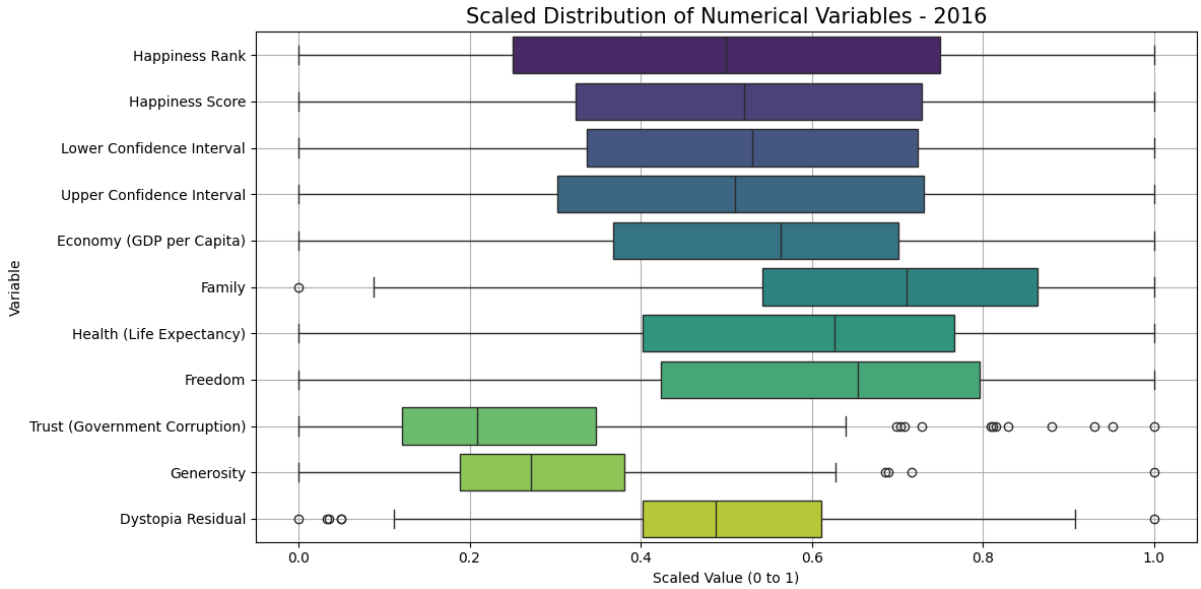
Year 2015
In 2015, a clear concentration around mid-to-high values was observed for variables such as *Happiness Score*, *Health (Life Expectancy)*, and *Family*, indicating their central role in shaping perceived well-being. Conversely, *Trust*

(Government Corruption) and *Generosity* displayed broader variability and numerous outliers, suggesting high heterogeneity across countries in terms of institutional trust and philanthropic behavior.



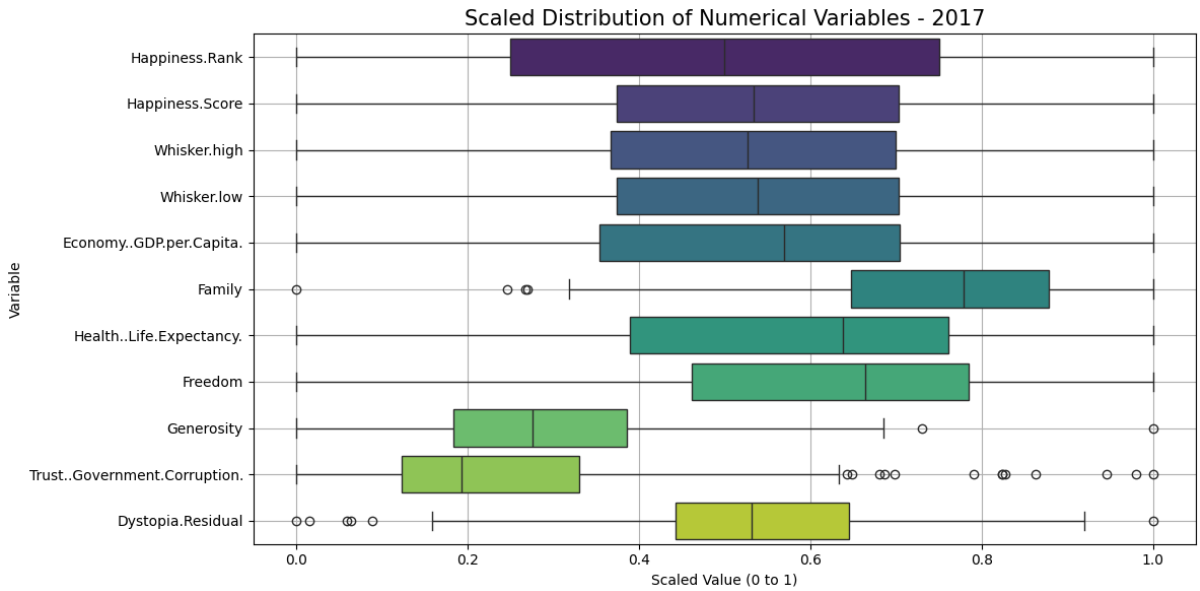
Year 2016

The 2016 distributions maintained patterns similar to the previous year. Variables such as *Health*, *Freedom*, and *Family* showed consistent interquartile ranges, reflecting cross-country stability. Newly included variables *Lower Confidence Interval* and *Upper Confidence Interval* revealed compact and symmetric distributions, reinforcing the statistical reliability of the happiness score estimates. As in 2015, *Trust* and *Generosity* remained highly dispersed.



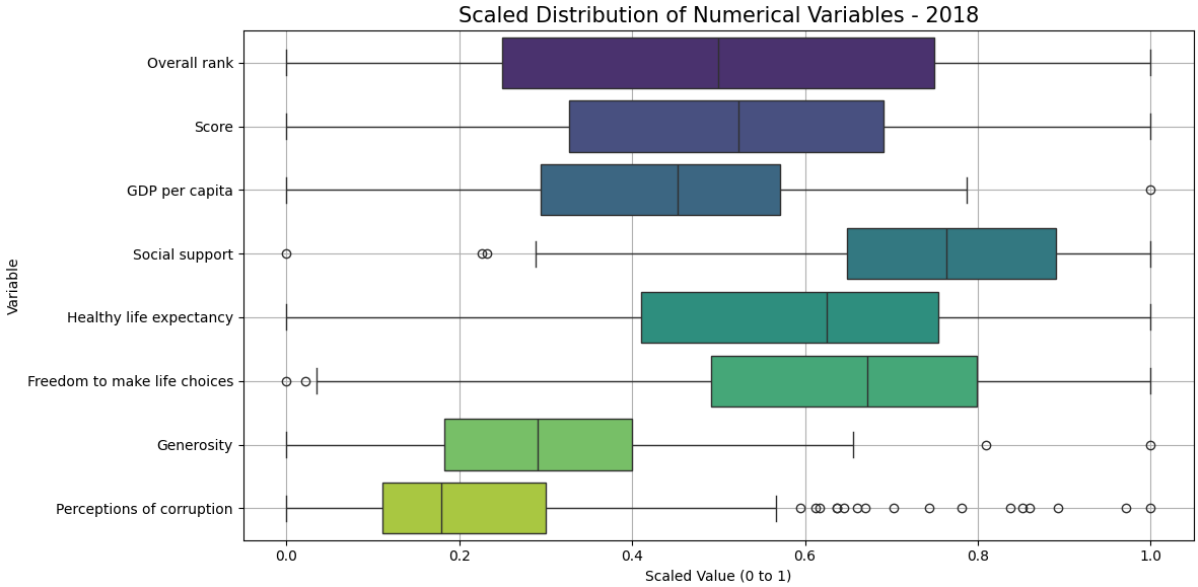
Year 2017

Key variables such as *Economy (GDP per Capita)*, *Health*, and *Family* continued to show stable and concentrated distributions. The introduction of *Whisker.high* and *Whisker.low* (confidence interval bounds) revealed symmetrical behavior, suggesting the robustness of predictions. However, *Trust (Government Corruption)* continued to exhibit dispersed and low values, underlining persistent institutional mistrust globally.



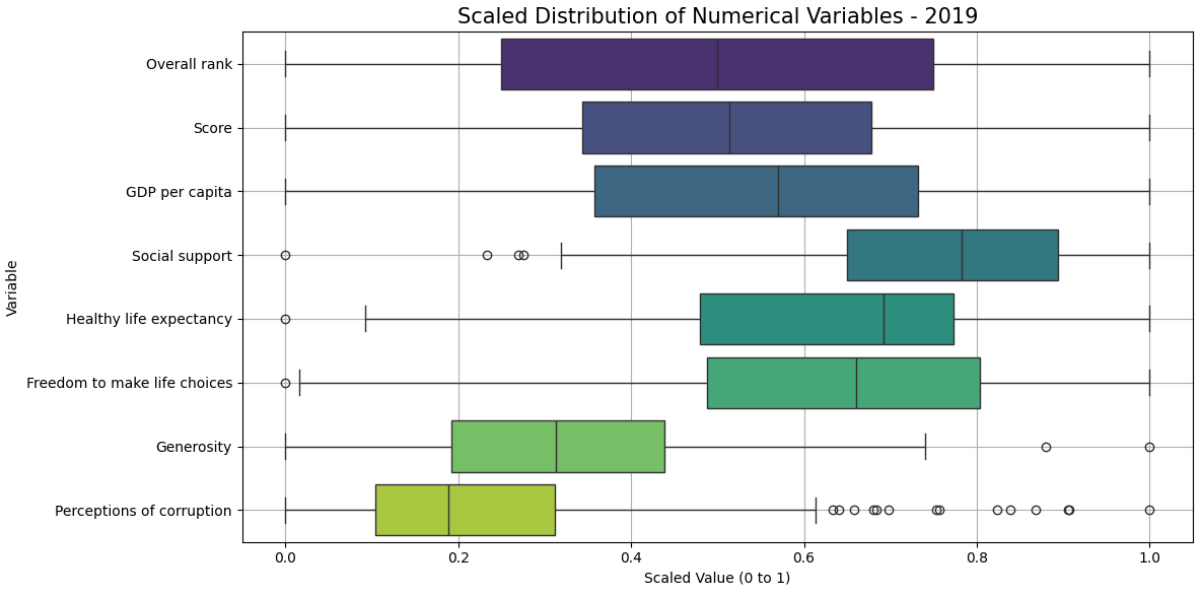
Year 2018

In 2018, *Healthy Life Expectancy* and *Social Support* again demonstrated distributional consistency, confirming their influence on happiness outcomes. Meanwhile, *Freedom to Make Life Choices* exhibited higher variability. As in prior years, *Generosity* and *Perceptions of Corruption* showed significant dispersion and outliers, highlighting structural and cultural differences between countries.



Year 2019

Finally, in 2019, a slight convergence trend was noted in the distributions of *Health*, *Freedom*, and *Social Support*, potentially indicating global alignment in these dimensions. Nonetheless, *Generosity* and *Perceptions of Corruption* retained wide spreads, underscoring persistent disparities. *Overall Rank* followed a near-normal distribution, suggesting that the model captured the underlying structure of well-being rankings effectively.



6. Model Training and Performance Evaluation

The modeling phase aimed to construct an accurate regression system capable of predicting the *Happiness Score* from a series of socioeconomic and psychological indicators. Three supervised regression models were trained and evaluated: Linear Regression, Ridge Regression, and Huber Regressor. All models were trained on a cleaned and standardized dataset covering the years 2015 to 2019.

6.1 Model Selection

The choice of models was guided by their robustness, interpretability, and ability to manage outliers:

- **Linear Regression** served as the baseline due to its simplicity and transparency.
- **Ridge Regression**, a regularized variant, was selected to address potential multicollinearity.
- **Huber Regressor** was included for its robustness to outliers, balancing squared and absolute error loss.

All models were trained using a 70/30 train-test split, and feature scaling was applied using StandardScaler. The models were saved using joblib for later deployment in the Kafka consumer pipeline.

6.2 Performance Metrics

The evaluation included several metrics:

- **R^2 Score (coefficient of determination)**
- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **Prediction Accuracy within a ±0.3 tolerance**

Below are the results obtained:

Model	R^2	RMSE	MAE	MSE	Acurracy
Linear Regresion	**0.9863**	**0.1307**	**0.09**	**0.017**	**97.02%**
Ridge Regresion	**0.9862**	**0.1313**	**0.0986**	**0.0172**	**97.02%**
Huber Regressor	**0.9859**	**0.1328**	**0.0938**	**0.0176**	**96.60%**

6.3 Model Comparison Visualization

The following figures illustrate the relationship between actual and predicted happiness scores for each regression model:

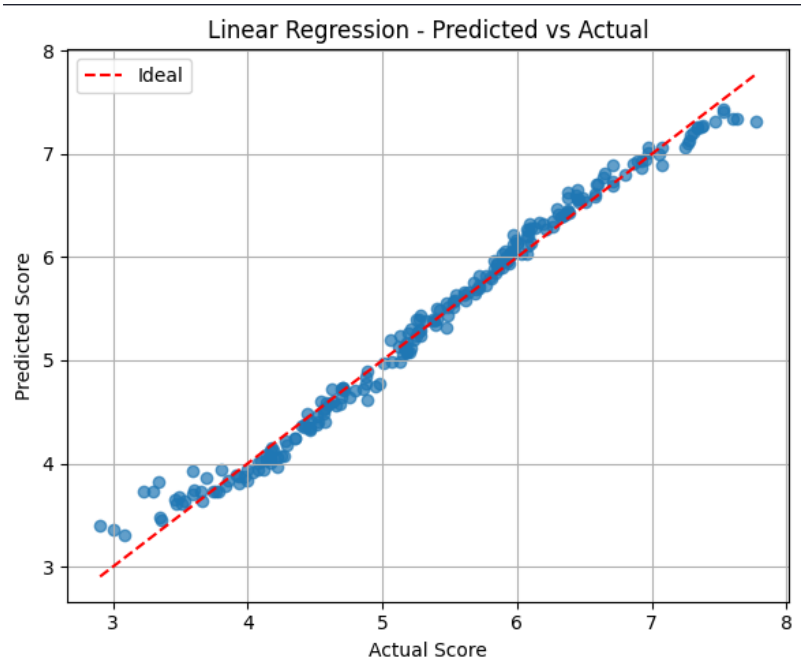


Figure 1. Linear Regression – Predicted vs Current:
The data points are tightly clustered along the red dashed line, indicating that the linear regression model achieved strong alignment between actual and predicted values. Minor deviations are visible in lower and upper score ranges, but overall accuracy is high.

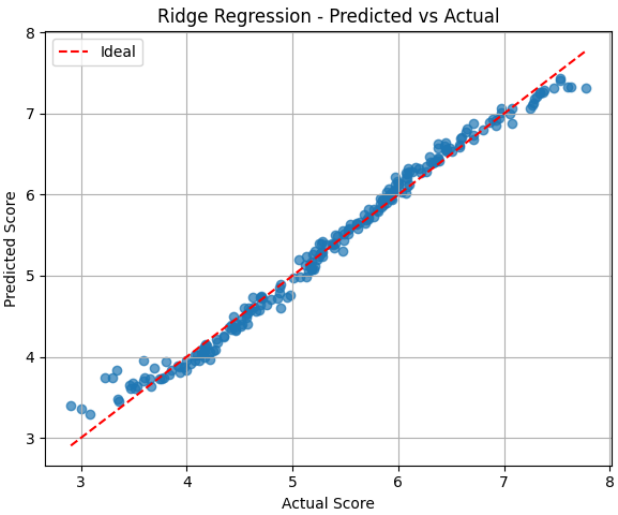


Figure 2. Ridge Regression – Predicted vs Current:
Ridge regression shows a very similar pattern to linear regression. The predicted scores follow the ideal line closely, with slightly better regularization visible in extreme values. This model effectively reduces the potential overfitting of the standard linear model.

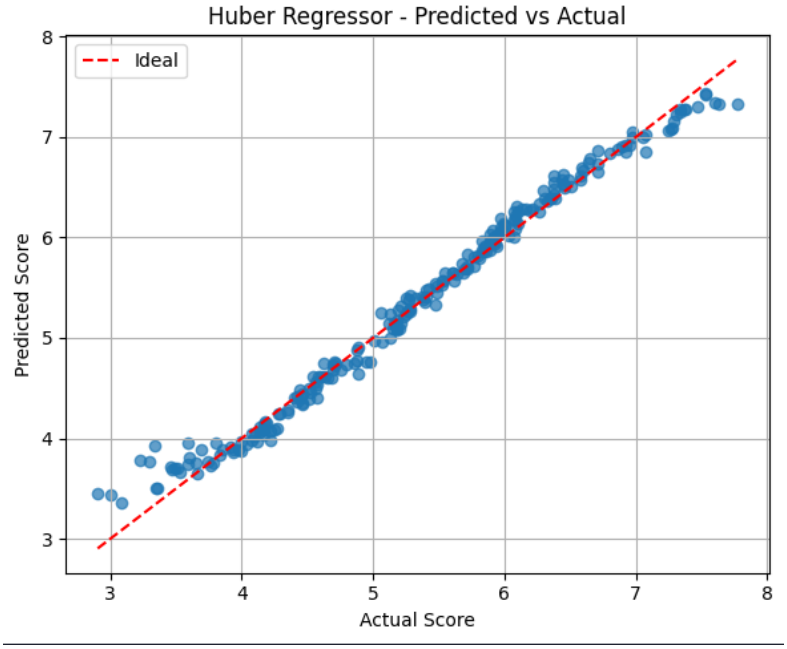


Figure 3. Huber Regressor – Predicted vs Actual:
The Huber regressor also performs well and is particularly robust to outliers. The predicted scores align closely with actual values, although some dispersion is noticeable. It balances between the sensitivity of linear regression and the robustness of regularized models.

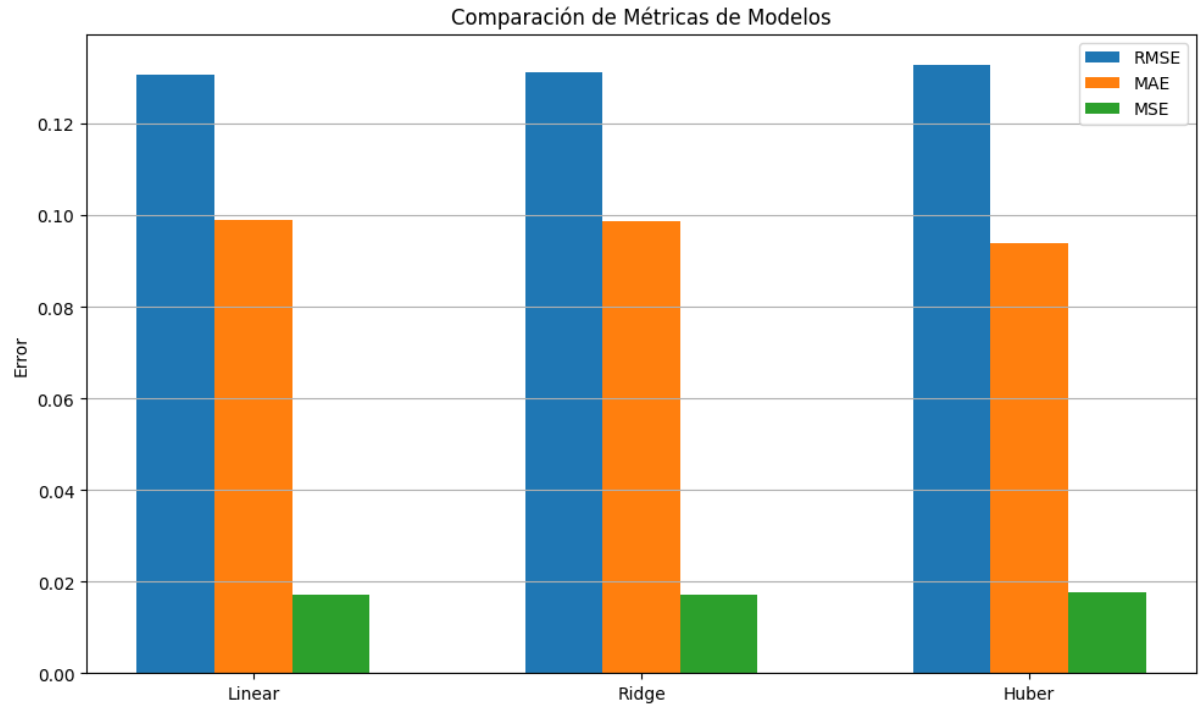


Figure 4. Comparison of Error Metrics Across Models:
This bar chart compares RMSE, MAE, and MSE across the three models. All models exhibit similar performance, with Linear Regression and Ridge Regression showing slightly lower RMSE and MSE. Huber Regressor, however, achieves the lowest MAE, suggesting it better handles small deviations.

6.4 Final Decision

Despite all models performing similarly, Linear Regression was selected as the final model. Its high accuracy, interpretability, and minimal computational complexity made it the most suitable for integration into the real-time pipeline.

7. Real-Time Inference and Data Streaming Deployment

To operationalize the trained model, a real-time data streaming architecture was implemented using **Apache Kafka** and **PostgreSQL**, allowing continuous ingestion, processing, prediction, and storage of happiness-related data.

7.1 Kafka-Based Streaming Pipeline

The streaming component is composed of two main scripts: a **Producer** and a **Consumer**.

- **Producer:**
  The Kafka producer reads rows from the cleaned dataset (happiness_cleaned.csv) and sends each record as a JSON message to the Kafka topic happiness_topic. The records exclude non-feature fields such as country names and original happiness scores, preserving only the input features required by the model.
- **Consumer:**
  The consumer subscribes to the happiness_topic, receives each message in real time, scales the features using the saved scaler.pkl, and applies the happiness_model.pkl to generate a predicted happiness score. The result is then stored along with the input features in a PostgreSQL database table named predictions.
  This streaming loop supports **online inference**, ensuring the pipeline can process and store predictions as new data arrives.

7.2 PostgreSQL Storage

The output from the consumer is stored in a **PostgreSQL** database named happiness_predictions. The schema includes both input variables (such as GDP, Freedom, Social support) and the corresponding model-generated predicted_score. This setup enables later querying, dashboard visualization, or batch analysis of predictions.
The table structure is defined as:

```
CREATE TABLE IF NOT EXISTS predictions (
    id SERIAL PRIMARY KEY,
    gdp FLOAT,
    social_support FLOAT,
    health FLOAT,
    freedom FLOAT,
```

```
    trust FLOAT,
    generosity FLOAT,
    dystopia FLOAT,
    year INT,
    predicted_score FLOAT
);
```

7.3 Containerized Execution with Docker

To ensure reproducibility and ease of deployment, all components—including Kafka, Zookeeper, PostgreSQL, and the inference application—were containerized using Docker Compose. By executing a single command, all services are launched automatically, enabling the full pipeline to run end-to-end without manual setup. This structure benefits both local testing and potential cloud deployment scenarios.

7.4 Monitoring Predictions

To validate the system's functionality, a query module (consultar_db.py) was implemented to fetch the most recent predictions from the PostgreSQL database. This allows rapid verification that the model is producing results correctly and that the pipeline is persisting records as expected.

8. Results and Final Conclusions

8.1 Summary of Results

After executing the entire data processing and machine learning pipeline, the following results were obtained:

- **The Linear Regression model** emerged as the most balanced and interpretable algorithm, achieving an R^2 score of 0.9863 and the lowest MAE (0.0990) among all models.
- **The Ridge Regression and Huber Regressor models** also performed well but offered no substantial improvements over the baseline linear model.
- All models demonstrated excellent alignment between actual and predicted happiness scores, as seen in the scatter plots, with most predictions falling very close to the ideal diagonal.

The implementation of real-time inference using Apache Kafka proved successful, allowing dynamic ingestion and scoring of new observations. Predictions were accurately stored in a PostgreSQL database, preserving both the features and model output for further visualization or auditing.

8.2 Conclusion

The project successfully combined Exploratory Data Analysis, feature transformation, model training, and streaming deployment into an integrated and automated pipeline. The use of containerization with Docker Compose simplified the orchestration of all services involved—Kafka, PostgreSQL, and the Python inference logic.

From a machine learning perspective, the insights gained from feature correlations, value distributions, and prediction accuracy confirmed the validity of the selected variables and the robustness of the linear model.
The pipeline, as deployed, is now capable of handling streaming data, generating real-time predictions, and persisting results in a scalable format. This architecture lays a solid foundation for further expansion, such as integrating a dashboard interface or enabling cloud-based deployments for real-world applications.