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), está 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
- **R² Score:** 0.7484799653866189

Ridge Regression for Regularization

- **Mean Squared Error (MSE):** 0.362724635227172
- **R² Score:** 0.719040811228492

Random Forest

- **Mean Squared Error (MSE):** 0.2706934794858266
- **R² Score:** 0.7903262888266669

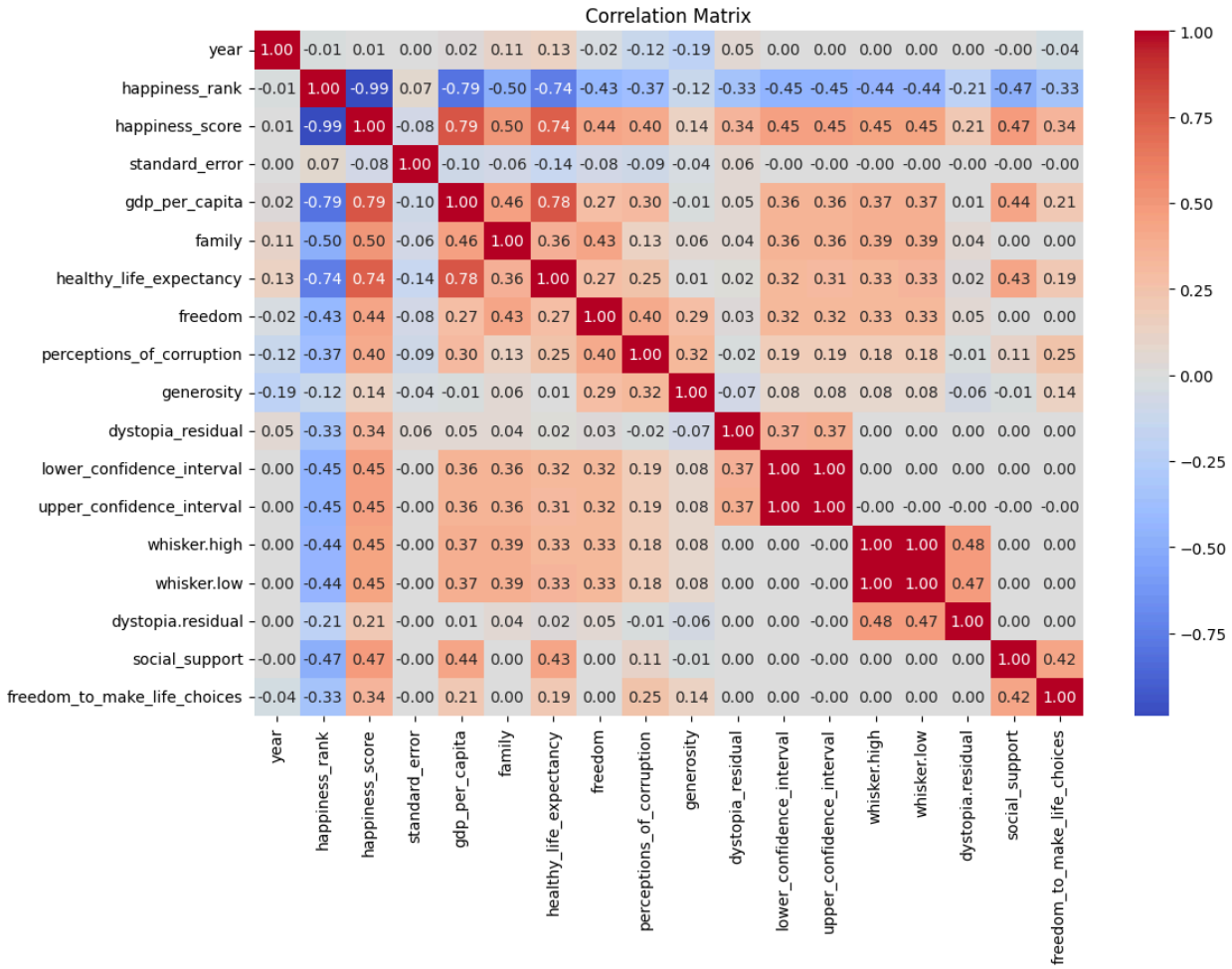
XGBoost

- **Mean Squared Error (MSE):** 0.2934685682195784
- **R² Score:** 0.7726851643112976

Stacking Regressor

- **Mean Squared Error (MSE):** 0.26137725175147697
- **R² Score:** 0.7975424509850881

El **mejor modelo** fue el Stacking Regressor con un **Mean Squared Error (MSE)** de 0.26137725175147697 y un **R² Score** de 0.7975424509850881.

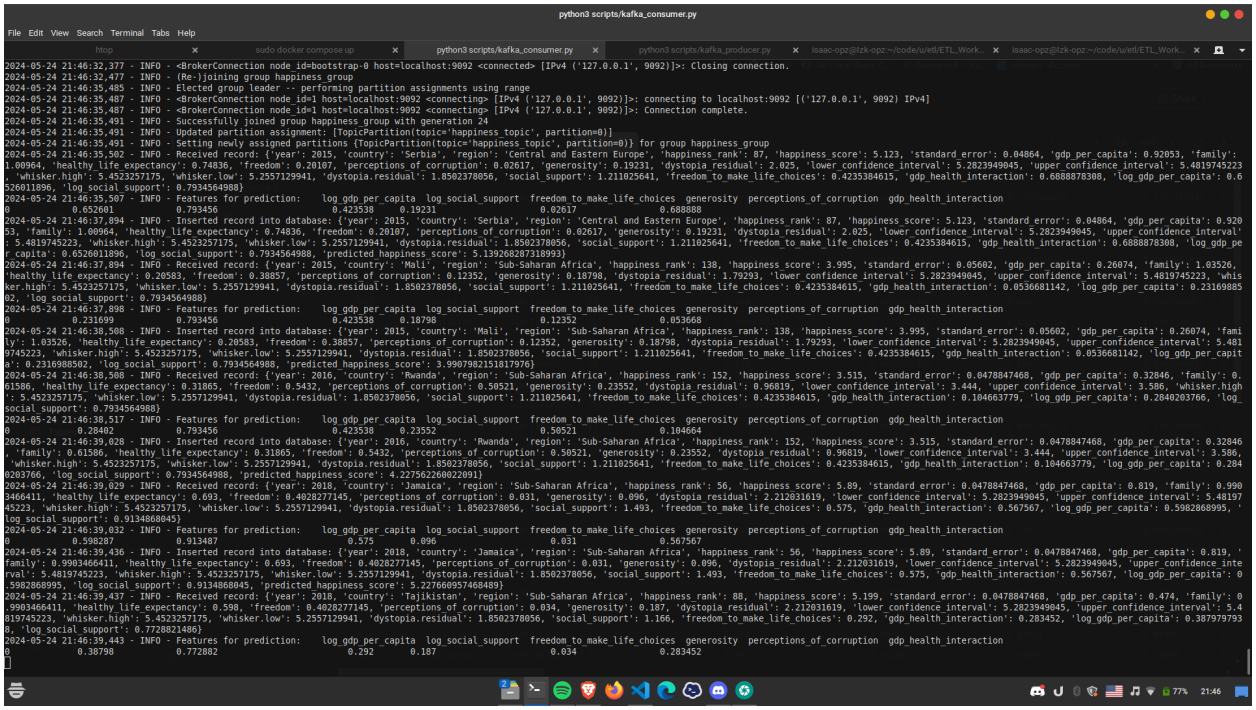


Matriz de correlación de los datos fuente

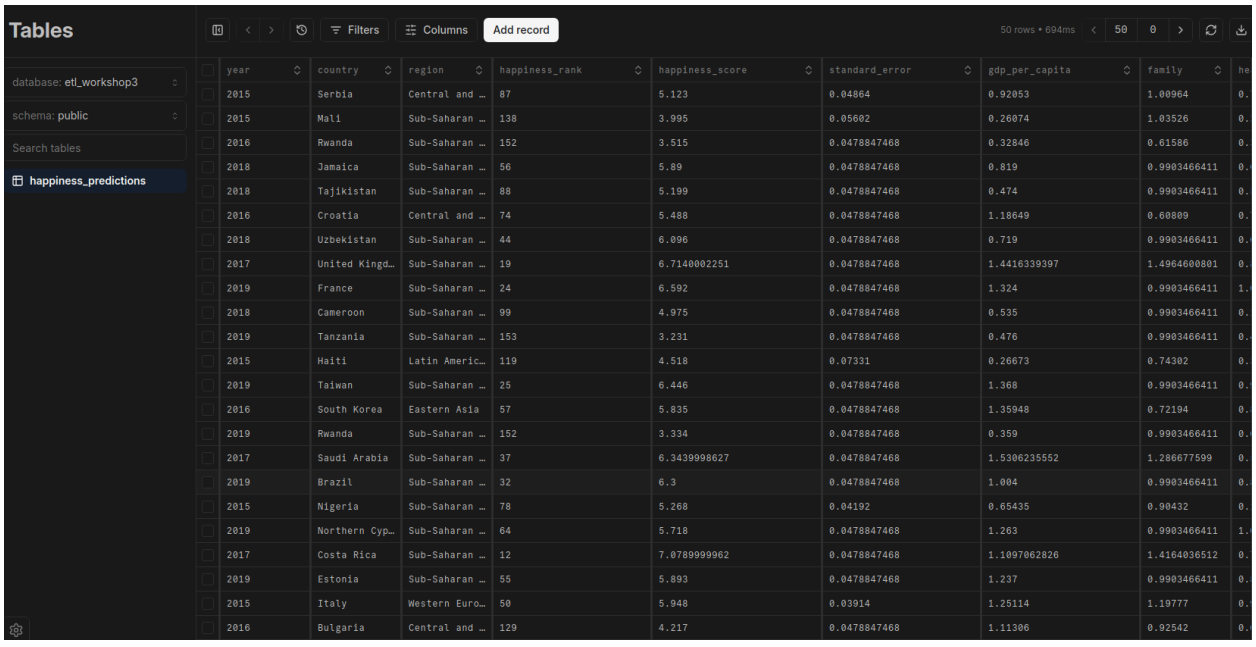
```
[21:47:03] isaac-opz ~/code/u/etl/ETL_Workshop3_Challenge
$ python3 scripts/performance_metric.py
/home/isaac-opz/.local/lib/python3.10/site-packages/scipy/
  warnings.warn(f"A NumPy version >={np_minversion} and <{
Data loaded from database:
   year     country                          region  happiness
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)



kafka consumer recibiendo datos y enviándolos a la base de datos



year	country	region	happiness_rank	happiness_score	standard_error	gdp_per_capita	family	he
2015	Serbia	Central and …	87	5.123	0.04864	0.92053	1.00964	0.
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.
2018	Jamaica	Sub-Saharan …	56	5.89	0.0478847468	0.819	0.9903466411	0.
2018	Tajikistan	Sub-Saharan …	88	5.199	0.0478847468	0.474	0.9903466411	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	0.9903466411	0.
2017	United Kingd…	Sub-Saharan …	19	6.7140002251	0.0478847468	1.4416339397	1.4964600801	0.
2019	France	Sub-Saharan …	24	6.592	0.0478847468	1.324	0.9903466411	1.
2018	Cameroon	Sub-Saharan …	99	4.975	0.0478847468	0.535	0.9903466411	0.
2019	Tanzania	Sub-Saharan …	153	3.231	0.0478847468	0.476	0.9903466411	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	0.9903466411	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	0.9903466411	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	0.9903466411	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	0.9903466411	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	0.9903466411	0.
2015	Italy	Western Euro…	50	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