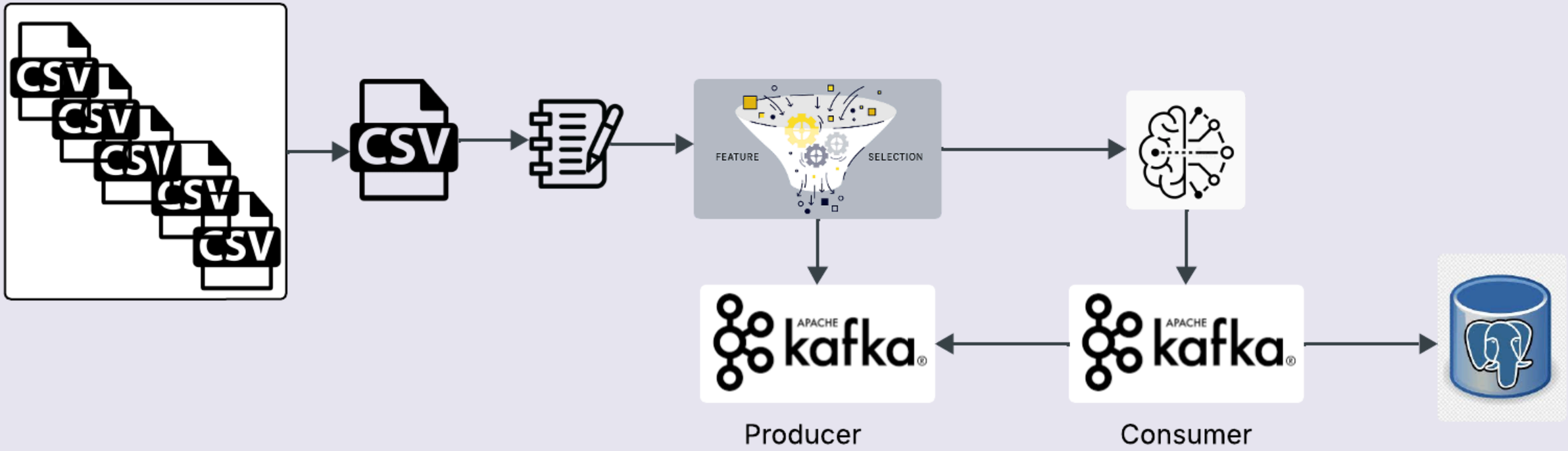
Prediccion puntajes de felicidad

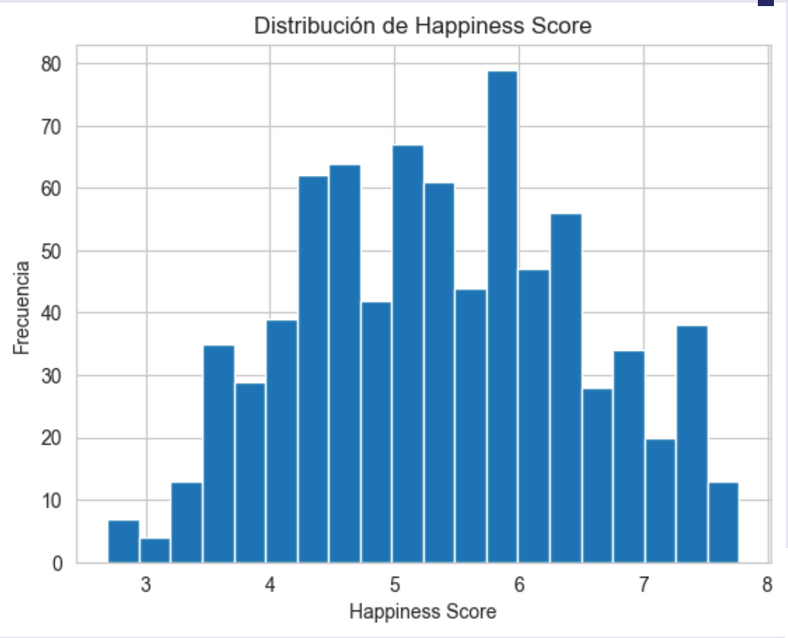
Machine Learning and Data Streaming

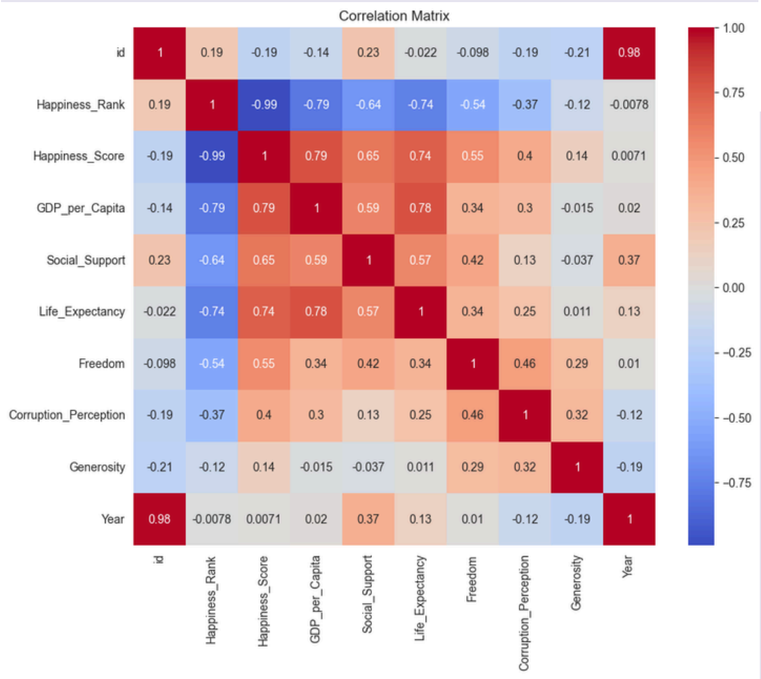
Presentado por Isabella Pérez Cav

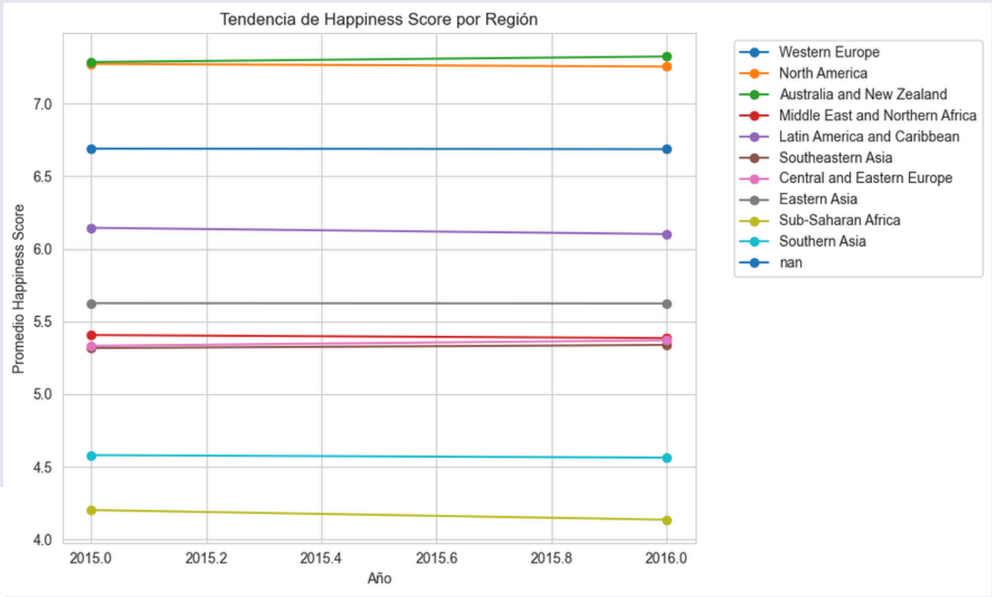
Pipeline


CSV
CSV
CSV
CSV
CSV
CSV
FEATURE
SELECTION
APACHE kafka
APACHE kafka
Producer
Consumer

Insights del eda







Preparacion datos

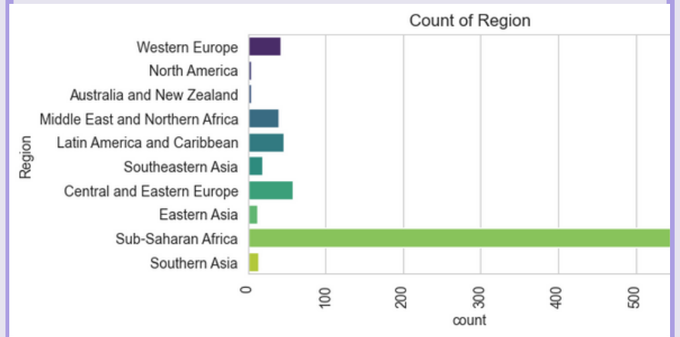
Imputar nulos

- Eliminar columnas con >50% de valores faltantes
- Imputamos con la moda global (Región) y rellenamos con la mediana(Corrupción)

	MissingCount
id	0
Country	0
Region	0
Happiness_Rank	0
Happiness_Score	0
GDP_per_Capita	0
Social_Support	0
Life_Expectancy	0
Freedom	0
Corruption_Perception	0
Generosity	0
Year	0

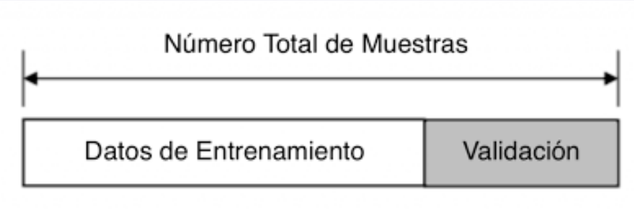
Transformación características

- Variables categóricas: codificación one-hot
- Variables numéricas: escalado estándar (StandardScaler)



División 80-20

- Los datos preprocesados se dividieron en:
  - 80% entrenamiento
  - 20% prueba



Features

- gdp_per_capita
- social_support
- life_expectancy
- freedom
- corruption_perception
- generosity
- Country_...

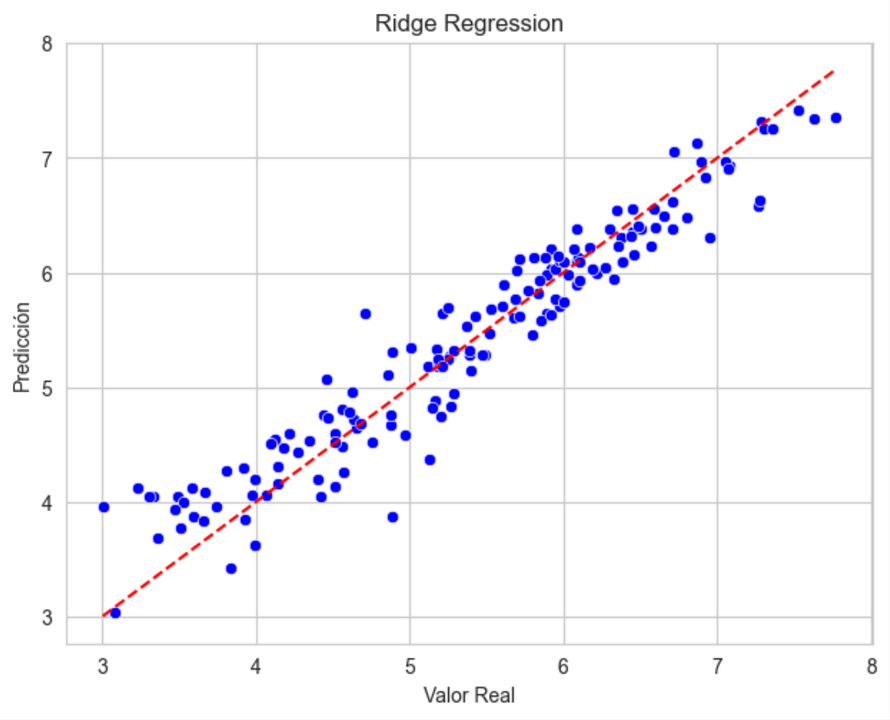
Se eliminaron:

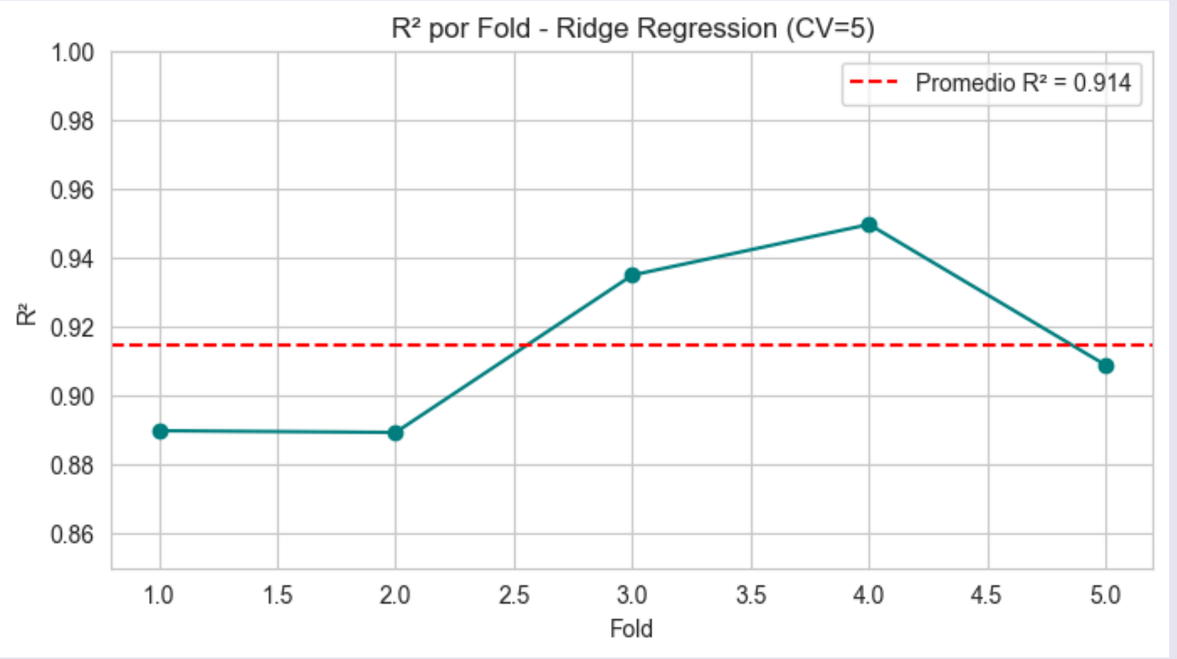
- Happiness_Score
- Region
- id
- Happiness_Rank
- Year
- Whisker_High
- Whisker_Low
- Upper_Confidence_Interval
- Lower_Confidence_Interval
- Standard_Error
- Dystopia_Residual

Comparacion resultados modelos

Modelo	MAE	RMSE	R^2
Ridge	0.243188	0.315927	0.917945
XGBoost	0.309364	0.396228	0.870931
Random Forest	0.350952	0.454441	0.830220
lightgbm	0.409087	0.519349	0.778257
Lasso	0.459663	0.585803	0.717880

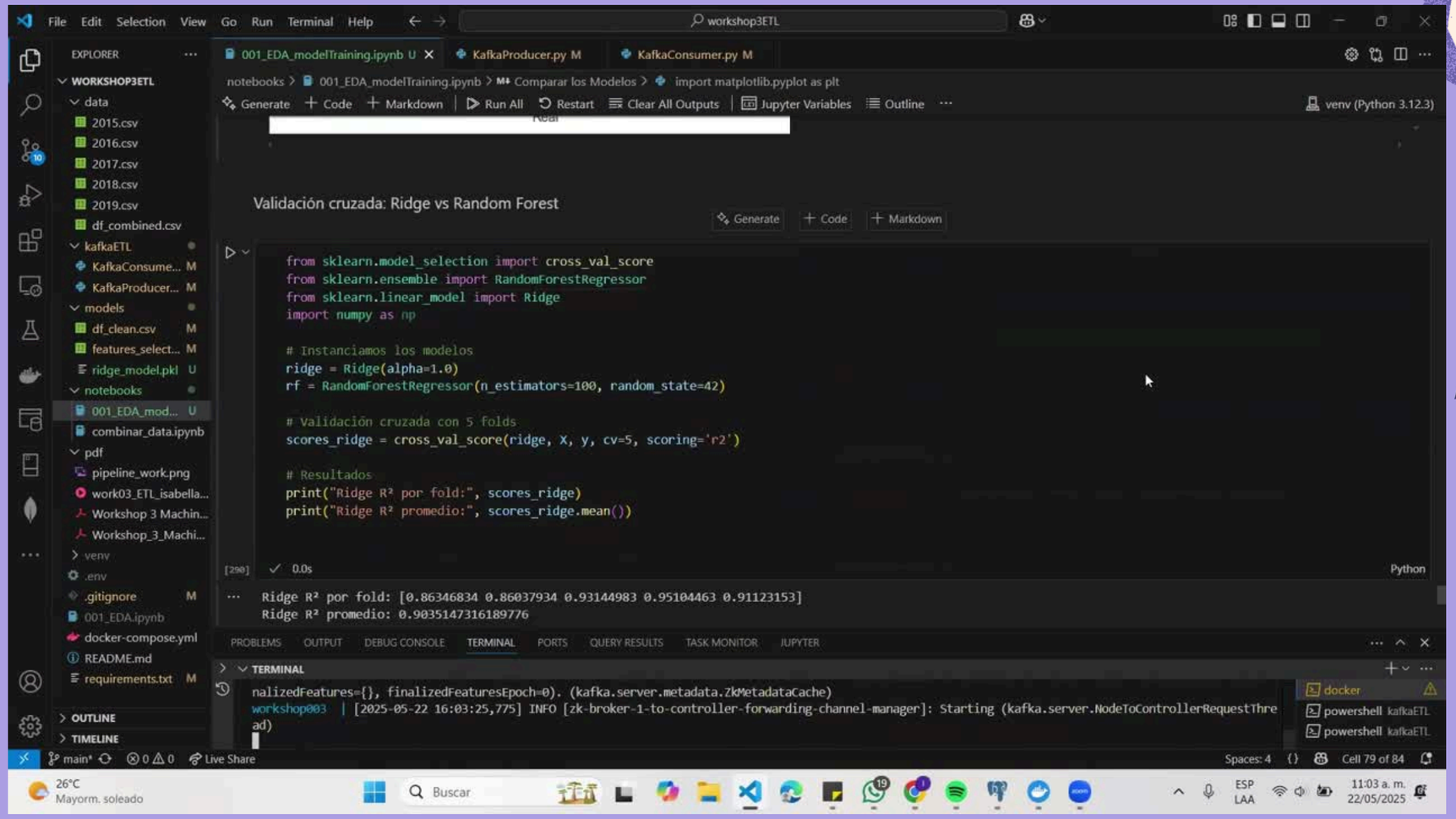
Validacion Ridge





Fold	R² (Ridge)
1	88.979.166
2	88.930.413
3	93.490.690
4	94.961.193
5	90.879.702
Promedio	91.448.233

Streaming con Kafka



Validacion modelo en la db

Valor Real = Happiness_Score

```
try:
    valor_real = float(data.get("valor_real"))
    acierto = abs(valor_real - prediction) < 0.1
except (TypeError, ValueError):
    valor_real, acierto = None, None
```

happiness_score_pred double precision	valor_real double precision	acierto boolean
6.156965558295278	6.455	false
6.297625400302368	6.411	false
6.290562056218954	6.329	true
6.458837020631479	6.302	false
6.188533872077306	6.298	true
6.354506702842515	6.295	true
6.04036599119472	6.269	false
6.123295656959879	6.168	true
5.902476571001644	6.13	false
6.0728443281362985	6.123	true
5.847482624771582	6.003	false
5.982434589431467	5.995	true
6.047628410480482	5.987	true
5.916206432593771	5.984	true
5.832115203292454	5.975	true
6.029597653557303	5.96	true
6.037082602448153	5.948	true

¡Muchas gracias!

Ya denle el titulo

Presentado por Isabella Pérez Cav