



Universidad Técnica Particular de Loja

PROYECTO BIMESTRAL

Autores

- Renata Maldonado
- Italo López
- Iván González
- Oliver Saraguro

**FUNDAMENTOS
DE
ANALISIS DE DATOS**

Docente:
Ing, Eduardo Encalada

OBJETIVO

Analizar un conjunto de datos sobre países de América Latina que contiene información geográfica, económica y social, incluyendo variables como población, PIB, desempleo, salario mínimo, y diversas categorías de indicadores.

El objetivo es clasificar las variables, realizar limpieza y transformación de los datos, identificar patrones a través de análisis exploratorio y aplicar técnicas de análisis predictivo multivariado (regresión y clasificación), con el fin de comprender mejor las dinámicas socioeconómicas de la región y generar visualizaciones complementarias en Power BI.





CLASIFICACION DE VARIABLES

CATEGÓRICAS

- continente
- regionPais
- subregionPais
- paisIndependiente
- tieneMar
- ONU
- anio
- decada
- unidad
- indicador
- dimension
- categoria
- nombrePais

CUANTITATIVAS

- latitud
- longitud
- areaPais
- sueldoMinimo_USD
- salarioMinimoPorHora_USD,
- valorIndicador
- pib_per_capita
- poblacion
- pib_total
- desempleo

BINARIAS

- paisIndependiente
- tieneMar
- ONU

TEXTO / ID

- bandera
- codigo_ISO
- capitalPais

ANALISIS APLICADOS

REGRESIÓN LINEAL MÚLTIPLE

Se utilizó como modelo base para predecir valorIndicador a partir de variables económicas y demográficas. Es útil para evaluar relaciones lineales entre variables.

ÁRBOL DE DECISIÓN (REGRESIÓN)

Aplicado para predecir valorIndicador de forma visual e interpretable. Permite entender cómo se dividen los datos según las variables más influyentes.

RANDOM FOREST (REGRESIÓN)

Se usó para mejorar la precisión de la predicción combinando múltiples árboles. Además, permite evaluar la importancia de cada variable.

ANALISIS APLICADOS

NAIVE BAYES (CLASIFICACIÓN)

Modelo de clasificación utilizado para predecir el continente al que pertenece un país, basándose en sus características. Rápido y eficaz con datos categóricos.

XGBOOST (REGRESIÓN)

Modelo avanzado de boosting utilizado por su alta precisión en predicciones. Fue uno de los mejores modelos del proyecto, con menor error (RMSE y MAE).

SVR – SUPPORT VECTOR REGRESSION

Aplicado para modelar relaciones no lineales entre valorIndicador y las demás variables. También obtuvo buen rendimiento, aunque ligeramente inferior a XGBoost.

REGRESIÓN LINEAL MÚLTIPLE

Objetivo:

Predecir el valorIndicador en función de variables económicas como PIB, población, desempleo, etc.

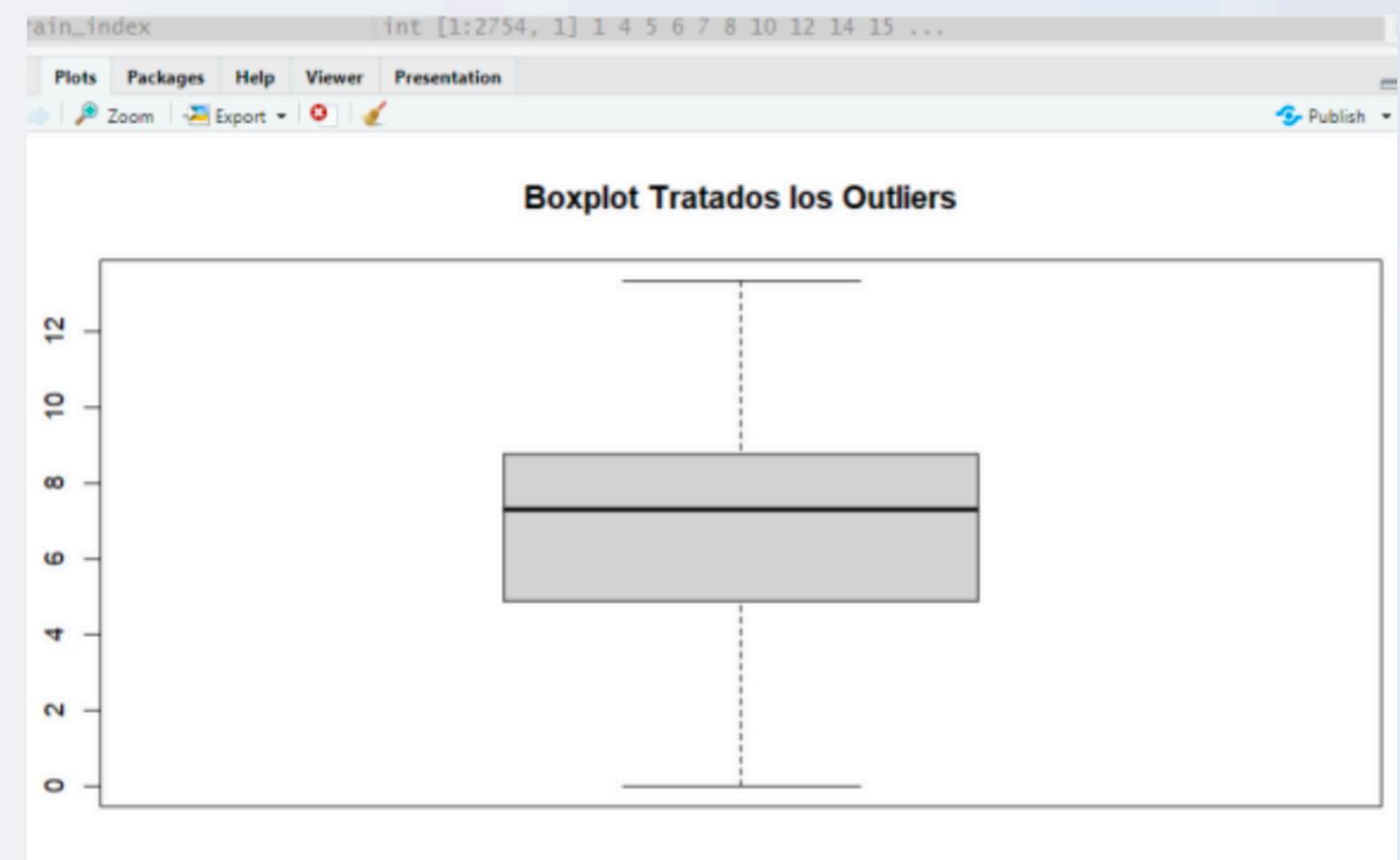
¿Por qué?

Sirve como modelo base para relaciones lineales y permite interpretar la influencia de cada variable.

Resultados:

- MAE: 1.47
- MSE: 4.17
- RMSE: 2.04

```
> # PREDICCIONES Y EVALUACIÓN
> # -----
> predictions <- predict(svr_model, newdata = test_data)
>
> mae <- mean(abs(predictions - test_data$valorIndicador))
> mse <- mean((predictions - test_data$valorIndicador)^2)
> rmse <- sqrt(mse)
>
> cat("MAE:", round(mae, 2), "\n")
MAE: 0.88
> cat("MSE:", round(mse, 2), "\n")
MSE: 2.03
> cat("RMSE:", round(rmse, 2), "\n")
RMSE: 1.42
```



MODELO ÁRBOL DE DECISIÓN - REGRESIÓN

Objetivo:

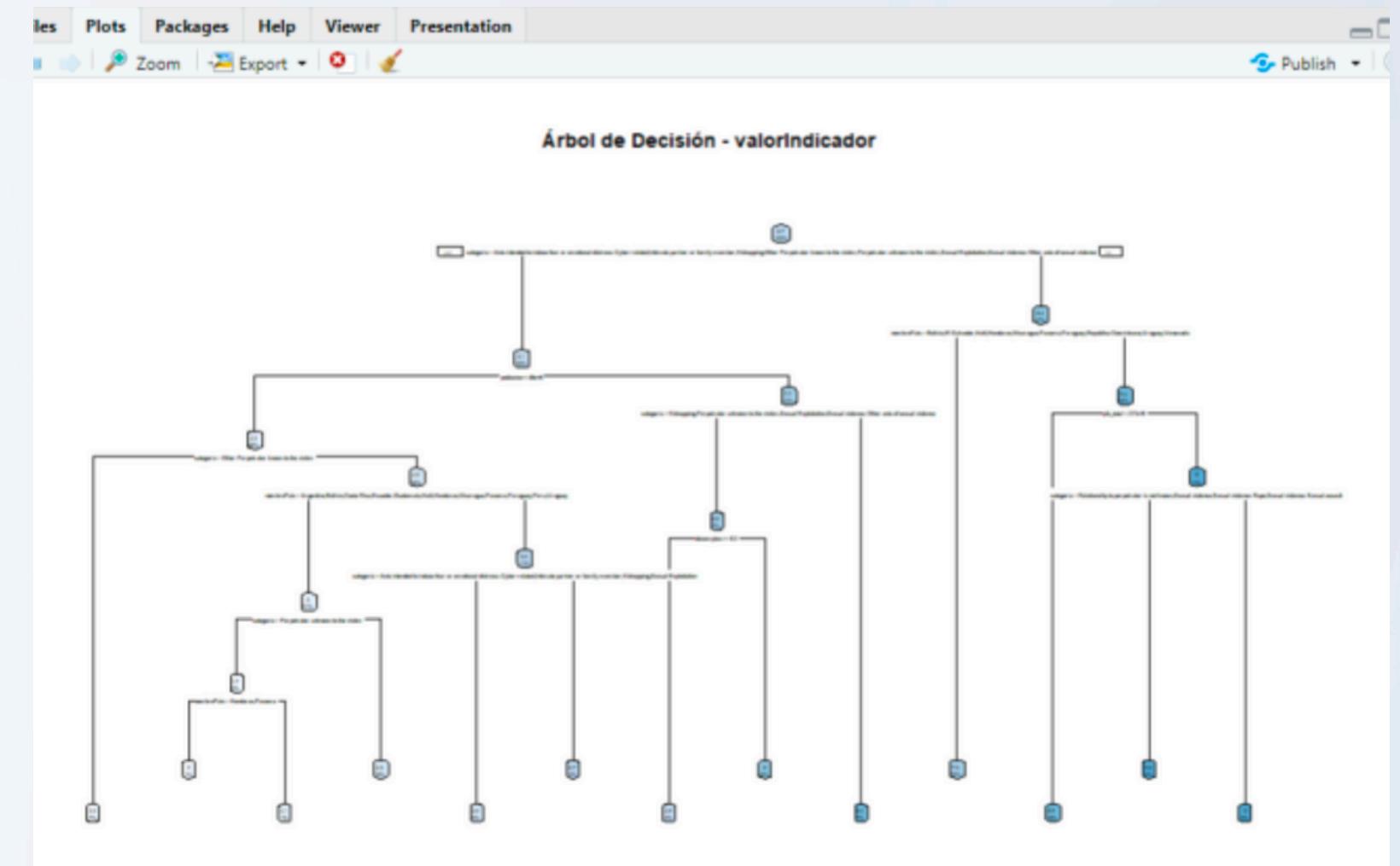
Predecir el valorIndicador con reglas interpretables.

¿Por qué?

Útil para visualizar cómo se dividen los datos según las variables clave.

Resultados:

- MAE: 1.34
- MSE: 3.04
- RMSE: 1.74



RANDOM FOREST (REGRESIÓN)

Objetivo:

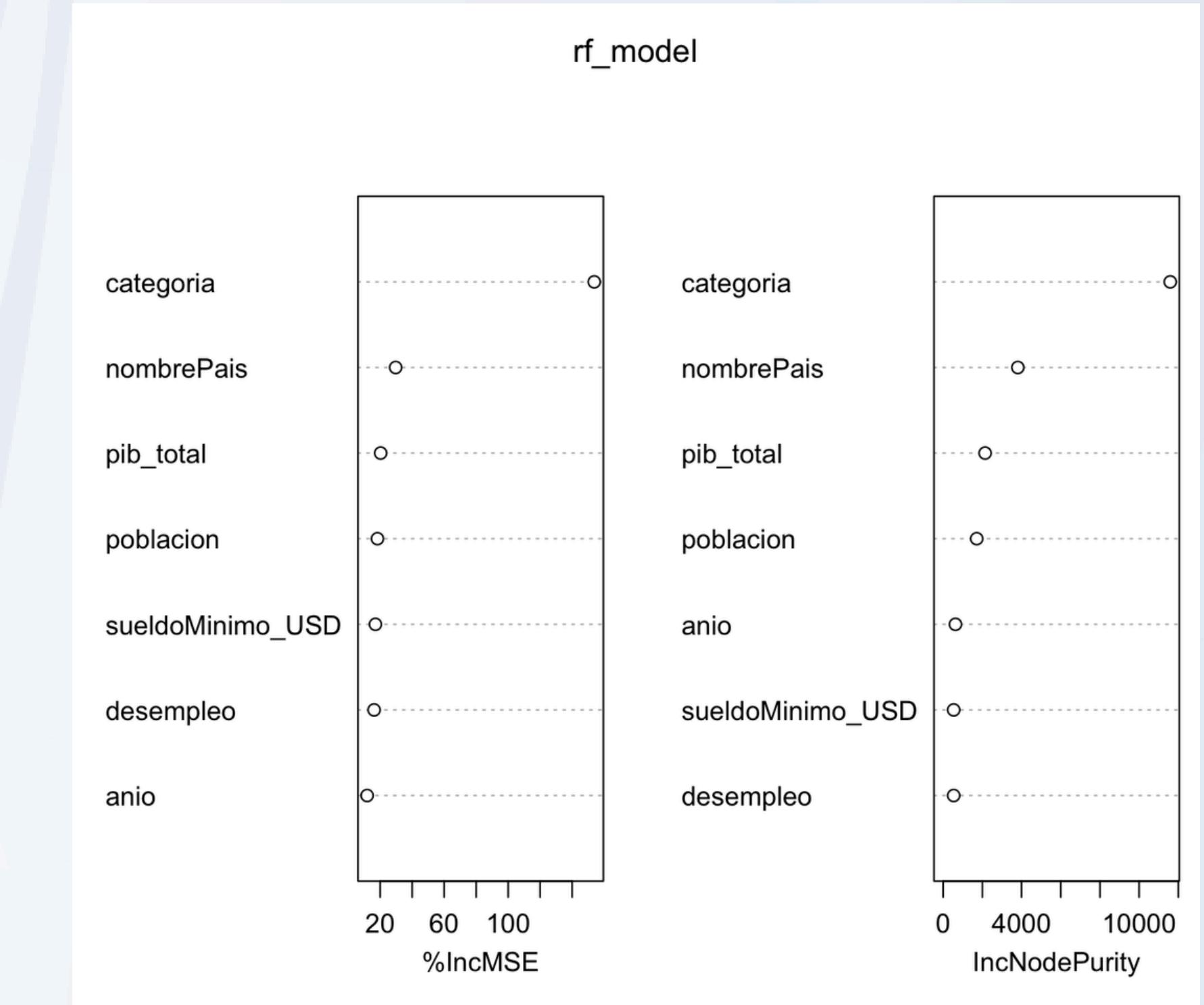
Mejorar la predicción del valorIndicador mediante el uso de múltiples árboles.

¿Por qué?

Reduce el sobreajuste y da mejor precisión que un solo árbol.

Resultados:

- MAE: 0.75
- MSE: 1.39
- RMSE: 1.18
- % Varianza explicada: 81.81%



MODELO NAIVE BAYES

- CLASIFICACIÓN

```

> # -----
> # Matriz de confusión
> # -----
> library(caret)
> conf_matrix <- confusionMatrix(nb_pred, test_data$continente)
> print(conf_matrix)

Confusion Matrix and Statistics

Reference
Prediction   North America South America
  North America      315          169
  South America       32          171

Accuracy : 0.7074
 95% CI : (0.6718, 0.7412)
No Information Rate : 0.5051
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4124

McNemar's Test P-Value : < 2.2e-16

  Sensitivity : 0.9078
  Specificity : 0.5029
  Pos Pred Value : 0.6508
  Neg Pred Value : 0.8424
  Prevalence : 0.5051
  Detection Rate : 0.4585
  Detection Prevalence : 0.7045
  Balanced Accuracy : 0.7054

'Positive' Class : North America

```

Objetivo:

Predecir el continente de un país usando variables económicas y sociales como PIB per cápita, desempleo, población, sueldos y categoría del indicador.

¿Por qué se aplicó este modelo?

Porque Naive Bayes es un clasificador eficiente para datos con muchas variables, especialmente mixtas. Nos permitió explorar si los indicadores económicos ayudan a distinguir países por región continental.

Resultados:

- Precisión del modelo: 70.7%
- Alta sensibilidad para países de América del Norte (90%)
- Menor especificidad para países de América del Sur (50%)
- El modelo mostró un buen desempeño general, aunque con margen de mejora al clasificar países del sur.

```

> # -----
> # EVALUACIÓN DEL MODELO
> # -----
> predictions <- predict(xgb_model, x_test)
> # Calcular métricas de error
> mae <- mean(abs(predictions - y_test))
> mse <- mean((predictions - y_test)^2)
> rmse <- sqrt(mse)
> cat("MAE:", round(mae, 2), "\n")
MAE: 0.69
> cat("MSE:", round(mse, 2), "\n")
MSE: 1.54
> cat("RMSE:", round(rmse, 2), "\n")
RMSE: 1.24

```

MODELO XGBOOST - REGRESIÓN

```
> # -----  
> # EVALUACIÓN DEL MODELO  
> # -----  
> predictions <- predict(xgb_model, x_test)  
> # Calcular métricas de error  
> mae <- mean(abs(predictions - y_test))  
> mse <- mean((predictions - y_test)^2)  
> rmse <- sqrt(mse)  
> cat("MAE:", round(mae, 2), "\n")  
MAE: 0.69  
> cat("MSE:", round(mse, 2), "\n")  
MSE: 1.54  
> cat("RMSE:", round(rmse, 2), "\n")  
RMSE: 1.24
```

Objetivo:

Predecir el valor del indicador usando variables socioeconómicas como país, sueldo mínimo, desempleo y PIB.

¿Por qué se usó?

XGBoost es potente para regresión con muchas variables, incluso categóricas codificadas.

Resultados:

- MAE: 0.69
- RMSE: 1.24
- Buen desempeño para estimar valores continuos.

MODELO DE REGRESIÓN SVR (SUPPORT VECTOR REGRESSION) - REGRESIÓN

Objetivo:

Predecir el valor del indicador utilizando variables como país, sueldo mínimo, desempleo, PIB y categoría del dato.

¿Por qué se aplicó?

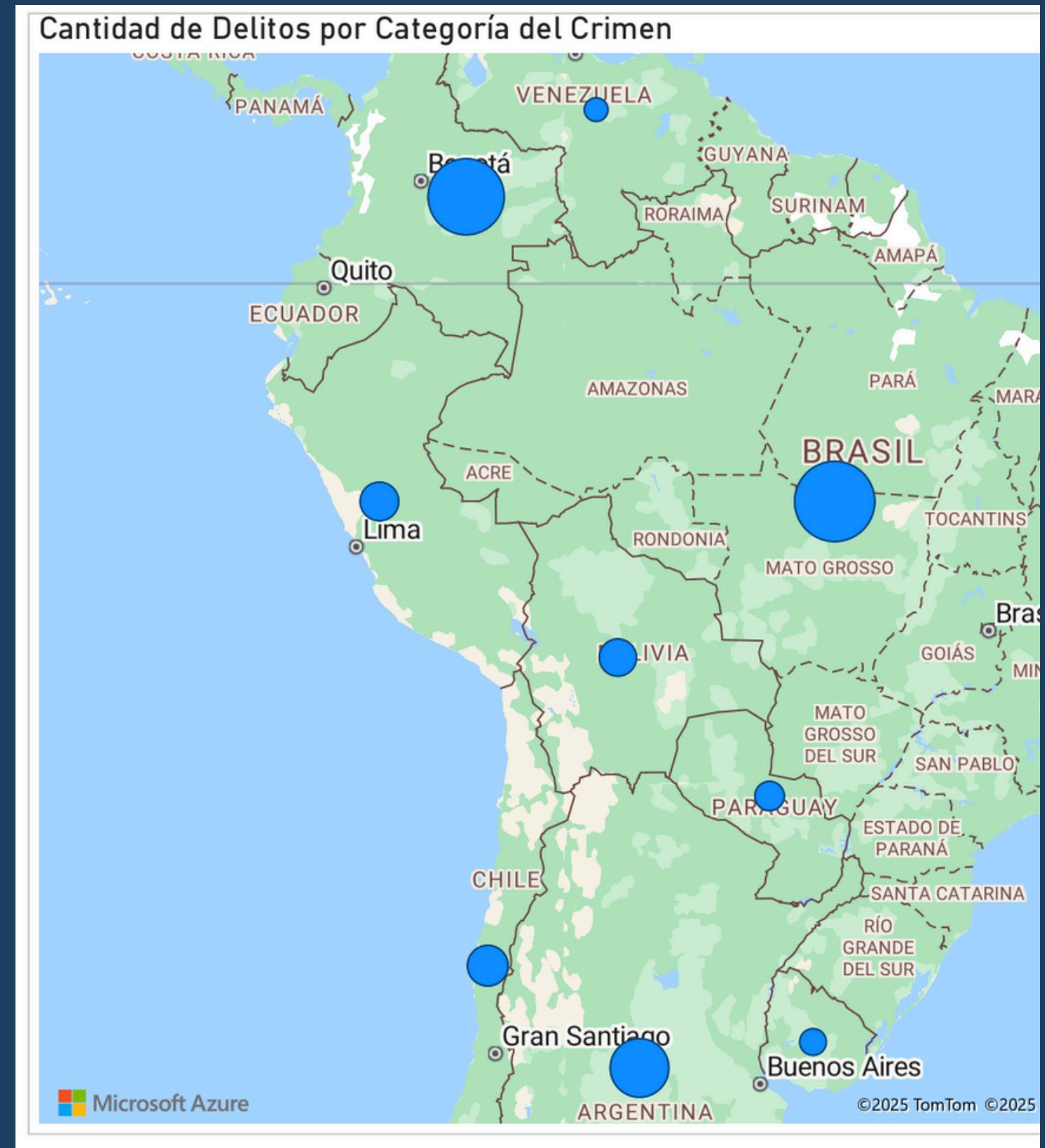
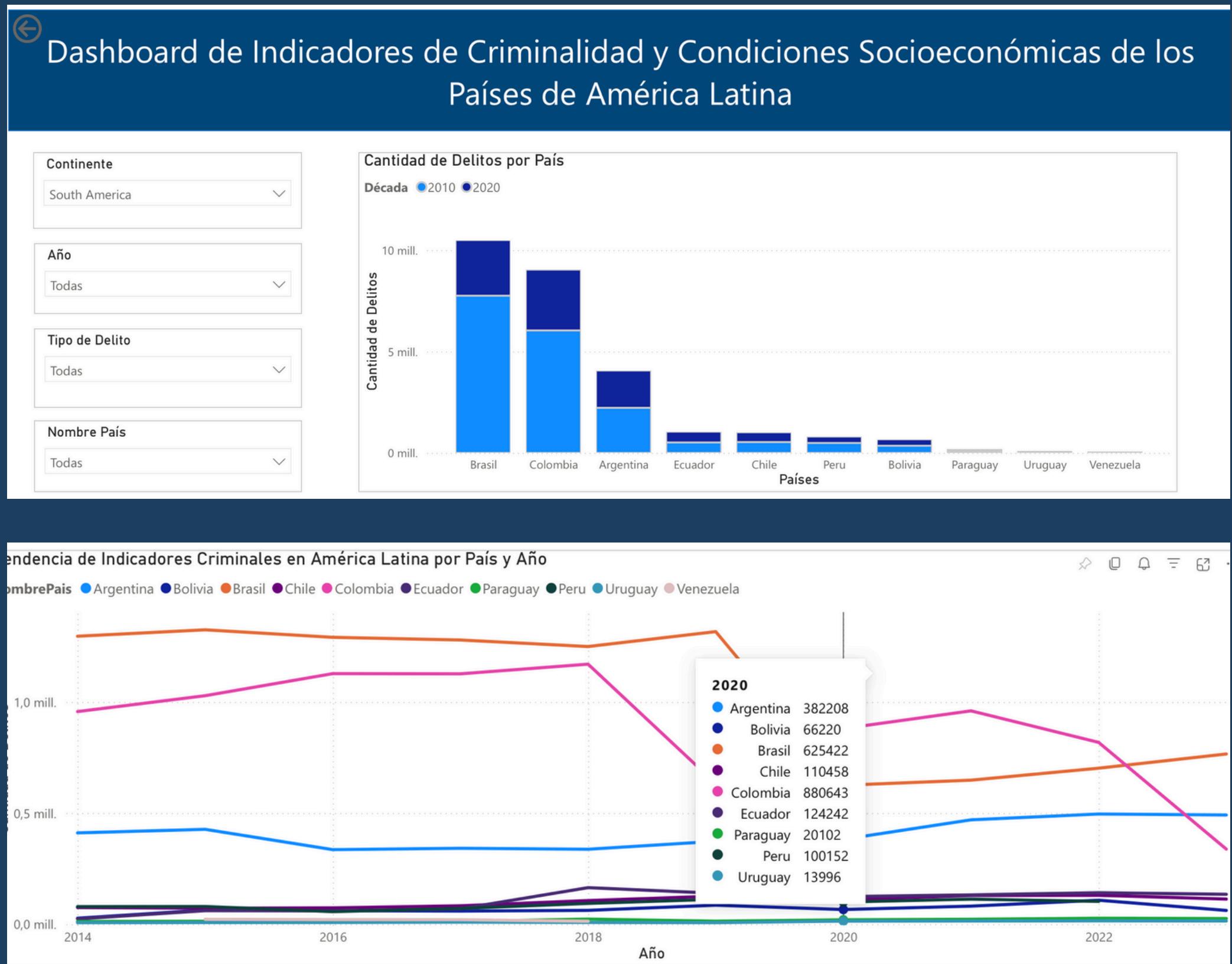
SVR permite capturar relaciones no lineales entre variables y ofrece buena generalización en regresión.

Resultados:

- MAE: 0.88
- RMSE: 1.42
- Desempeño aceptable, aunque con más error que otros modelos como XGBoost.

```
> # -----
> # PREDICCIONES Y EVALUACIÓN
> # -----
> predictions <- predict(svr_model, newdata = test_data)
> mae <- mean(abs(predictions - test_data$valorIndicador))
> mse <- mean((predictions - test_data$valorIndicador)^2)
> rmse <- sqrt(mse)
> cat("MAE:", round(mae, 2), "\n")
MAE: 0.88
> cat("MSE:", round(mse, 2), "\n")
MSE: 2.03
> cat("RMSE:", round(rmse, 2), "\n")
RMSE: 1.42
> |
```

POWER BI



CONCLUSION

A partir de un dataset sobre tasa de criminalidad en Latinoamérica, fuimos incorporando variables económicas y sociales que nos permitieron predecir y clasificar patrones delictivos con modelos como XGBoost y Naive Bayes. Los resultados muestran que estos factores influyen significativamente en los indicadores de criminalidad, y que es posible anticipar su comportamiento mediante técnicas de análisis predictivo.



Universidad Técnica Particular de Loja

MUCHAS
GRACIAS

Julio 2025