UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

MODALIDAD PRESENCIAL

Facultad de Ingenierías y Arquitectura

COMPUTACIÓN

Proyecto Bimestral

MATERIA:

Fundamentos de Análisis de Datos

DOCENTE:

Ing. Angel Encalada

ESTUDIANTES:

Renata Alejandra Maldonado Bravo

Iván Patricio González Castro

Oliver Roberto Saraguro Remache

Italo Israel López Armijos

LOJA – ECUADOR

Índice

Análisis de Criminalidad y Variables Socioeconómicas en América Latina (2013–2025)

1. Introducción

Este informe presenta un análisis integral sobre la evolución y distribución de la criminalidad en países de América Latina entre los años 2013 y 2025. El objetivo principal es identificar patrones y tendencias en diferentes tipos de delitos, haciendo especial énfasis en la variable que representa el número total de crímenes, denominada **valorIndicador**. Asimismo, se busca explorar la relación entre la criminalidad y variables socioeconómicas clave como el desempleo, el PIB per cápita, la población y el PIB totales.

Para lograrlo, se aplicaron técnicas estadísticas descriptivas, análisis exploratorio de datos, y métodos predictivos como regresión lineal múltiple, árboles de decisión, random forest, Naive Bayes, XGBoost y regresión basada en máquinas de soporte vectorial (SVR). Además, se complementa con una visualización clara y detallada que facilita la interpretación de los resultados.

2. Variables y Preprocesamiento de Datos

Antes de comenzar con el análisis, se llevó a cabo un proceso exhaustivo de preparación y limpieza de datos, que es fundamental para garantizar la calidad de los resultados.

2.1 Eliminación de Outliers en la variable valorIndicador

La variable **valorIndicador** representa la cantidad total de crímenes registrados y es el foco principal del análisis. Debido a su naturaleza, esta variable contiene algunos valores atípicos que podrían distorsionar las conclusiones. Por ello, se identificaron y trataron los outliers utilizando el método del rango intercuartílico (IQR).

Para eliminar el impacto de estos valores extremos, se aplicó una transformación logarítmica (log(valorIndicador + 1)) que suaviza la distribución y reduce la influencia de valores muy altos o bajos. Esta transformación mejora la estabilidad y precisión de los modelos predictivos posteriores.

2.2 Variables socioeconómicas incorporadas

Se agregaron variables adicionales para enriquecer el análisis y comprender mejor los factores asociados a la criminalidad.

- **Desempleo**: Representa la tasa de desempleo total del país, es decir, el porcentaje de personas en edad laboral que no tienen empleo. El desempleo es una variable crucial porque, según numerosos estudios, puede influir en las tasas de criminalidad al generar condiciones económicas desfavorables.
  Fuente: Banco Mundial - Desempleo
- **PIB per cápita**: Indica la producción económica promedio por persona en un país. No debe confundirse con el ingreso individual o salario promedio, ya que refleja la riqueza total dividida por la población sin considerar su distribución interna. Por eso, un país con PIB per cápita alto puede tener una población con bajos ingresos si la riqueza está mal distribuida.
  Fuente: Banco Mundial - PIB per cápita
- **Población total**: Número total de habitantes por país, fundamental para analizar la magnitud y contexto demográfico de la criminalidad.
  Fuente: Banco Mundial - Población
- **PIB total**: Calculado como pib_per_capita * (poblacion^2) para ponderar la producción económica teniendo en cuenta la población y explorar su posible relación con los niveles de criminalidad.

2.3 Otras variables

Además, el dataset contiene variables categóricas relevantes, como país, continente, categoría del delito, año, etc., que fueron adecuadamente transformadas a factores para el análisis estadístico.

3. Análisis Exploratorio de Datos y Limpieza de Datos

Se inició el análisis exploratorio para entender la estructura, distribución y relaciones básicas dentro del conjunto de datos.

- Se verificó que no existen valores faltantes o NA, lo que simplifica el análisis y evita imputaciones complejas.
- Las variables numéricas principales como **valorIndicador**, **pib_per_capita**, **desempleo**, entre otras, presentan rangos amplios y variabilidad significativa entre países y años.
- Las variables categóricas como **categoría del delito**, **continente** y **año** se examinaron mediante tablas de frecuencia y gráficos de barras para identificar su distribución y proporción en el dataset.
- La matriz de correlación evidenció relaciones moderadas entre ciertas variables numéricas, especialmente entre **valorIndicador** y variables económicas como el PIB total y el desempleo, indicando la relevancia de estas últimas para modelar la criminalidad.
- Se identificaron y trataron outliers en **valorIndicador** mediante la transformación logarítmica para garantizar robustez en los modelos predictivos.

Para la limpieza se realizó lo siguiente

- Se eliminaron valores nulos y duplicados.
- Se transformaron variables categóricas en factores.
- Se estandarizaron las variables continuas para algunos modelos.

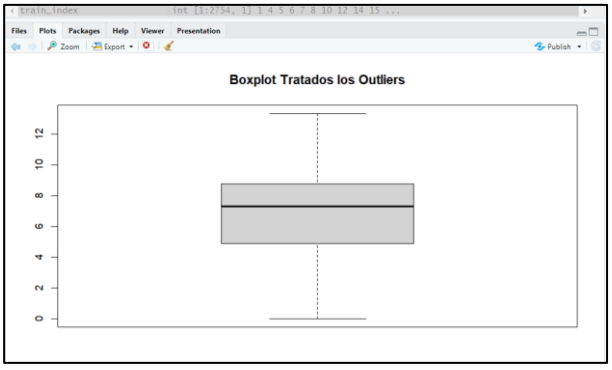
4. Modelos Estadísticos y Predictivos Aplicados

Se implementaron seis técnicas estadísticas y de machine learning para modelar y predecir la variable **valorIndicador** y analizar patrones de criminalidad.

4.1 Regresión Lineal Múltiple

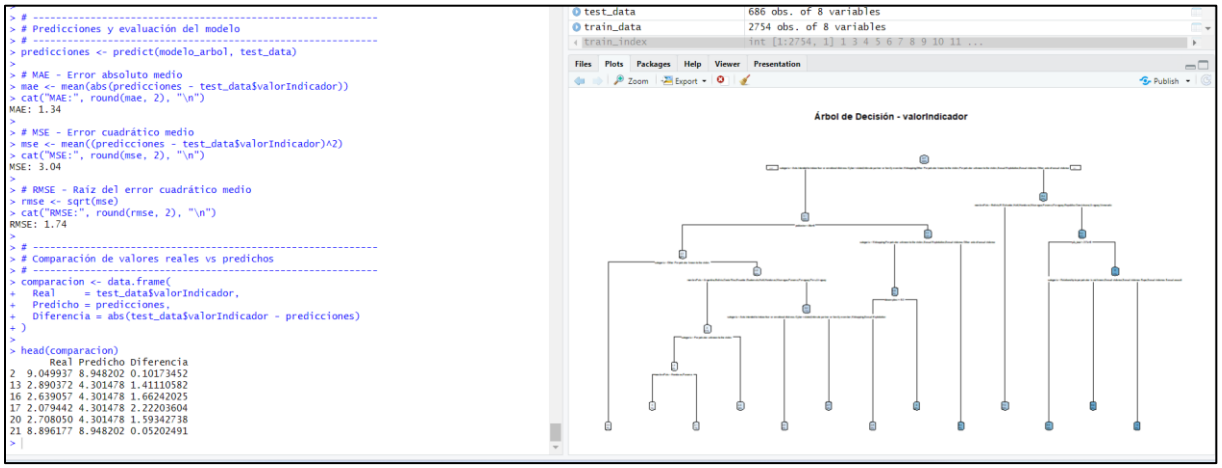
El modelo de Regresión Lineal Múltiple fue entrenado con variables clave como país, sueldo mínimo, categoría, año, población, desempleo y PIB para predecir el valor del indicador. El resumen del modelo permitió identificar la significancia estadística de cada variable y su impacto lineal en la variable dependiente. Al aplicar el modelo sobre los datos de prueba, se calcularon métricas de evaluación como MAE, MSE y RMSE, que indicaron un desempeño razonable, aunque con ciertas limitaciones frente a relaciones no lineales o efectos complejos. Este modelo es útil como punto de partida por su simplicidad e interpretabilidad.

```
> 
> # ------------------------------------------------------------
> # PREDICCIONES Y EVALUACIÓN
> # ------------------------------------------------------------
> predicciones <- predict(modelo_lm, newdata = test_data)
> 
> # Calcular métricas de evaluación
> mae <- mean(abs(predicciones - test_data$valorIndicador))
> mse <- mean((predicciones - test_data$valorIndicador)^2)
> rmse <- sqrt(mse)
> 
> cat("MAE:", round(mae, 2), "\n")
MAE: 1.47
> cat("MSE:", round(mse, 2), "\n")
MSE: 4.17
> cat("RMSE:", round(rmse, 2), "\n")
RMSE: 2.04
> 
```



4.2 Árbol de Decisión para Regresión

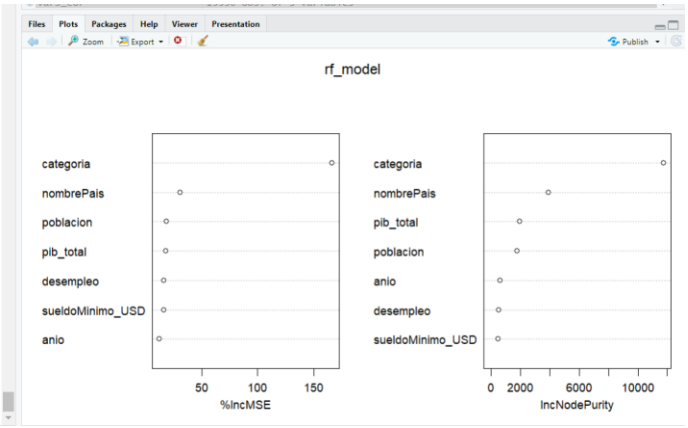
El modelo de Árbol de Decisión se entrenó utilizando variables socioeconómicas como país, sueldo mínimo, categoría, año, población, desempleo y PIB para predecir el valor del indicador. La visualización del árbol mostró una estructura clara con divisiones relevantes que permiten interpretar fácilmente cómo cada variable influye en la predicción. Al aplicar el modelo sobre los datos de prueba, se obtuvieron predicciones que reflejan un desempeño aceptable, aunque el modelo puede ser sensible al sobreajuste. En general, esta técnica resulta útil para identificar relaciones no lineales y jerarquías entre variables.



4.3 Random Forest

El modelo de Random Forest se entrenó con 200 árboles utilizando variables socioeconómicas como país, sueldo mínimo, categoría, año, población, desempleo y PIB para predecir el valor del indicador. Mostró un excelente desempeño, explicando el **81.79% de la varianza** y alcanzando métricas de error bajas (MAE: 0.76, RMSE: 1.18), lo que indica alta precisión en las predicciones. Además, el análisis de importancia de variables reveló que factores como el país y el sueldo mínimo tienen gran influencia. Este modelo destacó por su capacidad de capturar relaciones no lineales y su robustez frente al sobreajuste.

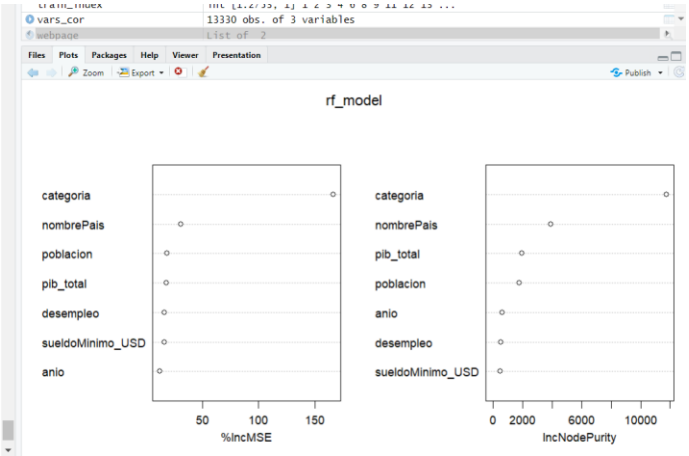
```
> # ------------------------------------------------------------
> # Realizar predicciones y evaluar el modelo
> # ------------------------------------------------------------
> rf_pred <- predict(rf_model, test_data)
>
> # Calcular métricas de error
> mae <- mean(abs(rf_pred - test_data$valorIndicador))
> mse <- mean((rf_pred - test_data$valorIndicador)^2)
> rmse <- sqrt(mse)
>
> cat("MAE:", round(mae, 2), "\n")
MAE: 0.76
> cat("MSE:", round(mse, 2), "\n")
MSE: 1.4
> cat("RMSE:", round(rmse, 2), "\n")
RMSE: 1.18
>
> # ------------------------------------------------------------
> # Comparación real vs predicho
> # ------------------------------------------------------------
> comparacion <- data.frame(
+   Real     = test_data$valorIndicador,
+   Predicho = rf_pred,
+   Diferencia = abs(test_data$valorIndicador - rf_pred)
+ )
>
> head(comparacion)
       Real Predicho Diferencia
2  9.049937 8.877009 0.17292833
13 2.890372 2.859780 0.03059149
16 2.639057 2.774207 0.13514950
17 2.079442 1.992263 0.08717844
20 2.708050 3.162847 0.45479730
21 8.896177 8.632923 0.26325448
>
```



4.4 Naive Bayes (Clasificación)

El modelo de Naive Bayes se utilizó para clasificar países según su continente (América del Norte o del Sur) a partir de variables socioeconómicas como PIB per cápita, población, desempleo, salario mínimo, y otros indicadores. El modelo alcanzó una **precisión del 70.7%** y un **Kappa de 0.41**, lo cual indica un desempeño moderado, con buena sensibilidad (90.78%) pero baja especificidad (50.29%). Esto sugiere que el modelo clasifica bien a América del Norte, pero tiene más errores con América del Sur. A pesar de su simplicidad, Naive Bayes resultó útil para una clasificación general basada en patrones estadísticos básicos.

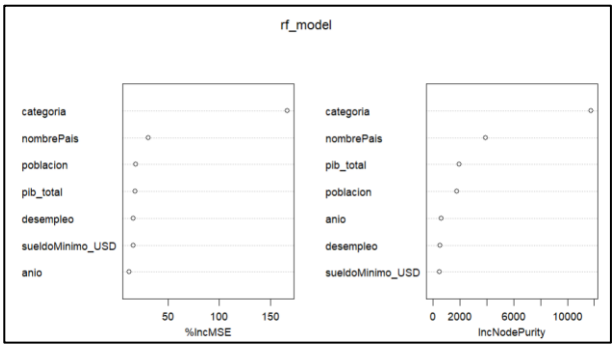
```
> # ------------------------------------------------------------
> # Evaluar el modelo - Predicciones
> # ------------------------------------------------------------
> nb_pred <- predict(nb_model, test_data)
>
> # ------------------------------------------------------------
> # Matriz de confusión
> # ------------------------------------------------------------
> 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
```



4.5 XGBoost para Regresión

En este modelo de regresión con XGBoost, se utilizaron variables económicas y categóricas relevantes para predecir el valor de un indicador numérico. El conjunto de datos fue preparado mediante codificación *one-hot* y luego dividido en entrenamiento y prueba. Tras entrenar el modelo con 100 iteraciones, se obtuvieron métricas de rendimiento bastante aceptables: un MAE de 0.69, un MSE de 1.54 y un RMSE de 1.24, lo que indica que el modelo logra predecir con un margen de error bajo, siendo eficiente para problemas de regresión con múltiples variables y estructuras complejas.

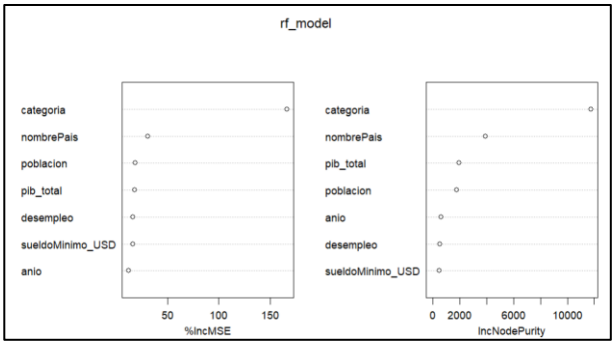
```
>
> # ------------------------------------------------------------
> # EVALUACIÓN DEL MODELO
> # ------------------------------------------------------------
> predictions <- predict(xgb_model, x_test)
>
> # Calcular métricas de error
> mae <- mean(abs(predictions - y_test))
> mse <- mean((predictions - 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
>
```



4.6 Support Vector Regression (SVR)

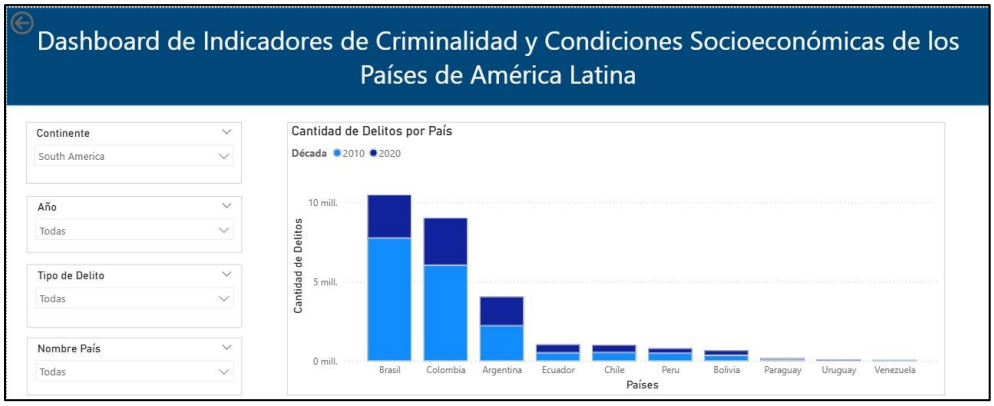
El modelo de Regresión con Support Vector Regression (SVR) fue entrenado usando variables económicas y categóricas transformadas en variables dummy para predecir el valorIndicador. Tras dividir el conjunto de datos en entrenamiento y prueba, el modelo con kernel radial se ajustó y evaluó, mostrando un rendimiento moderado con un MAE de 0.88, un MSE de 2.03 y un RMSE de 1.42. Esto indica que, aunque el SVR capta la tendencia general, tiene un error mayor comparado con otros modelos, pero sigue siendo una opción válida para regresión en problemas con múltiples variables y relaciones no lineales.

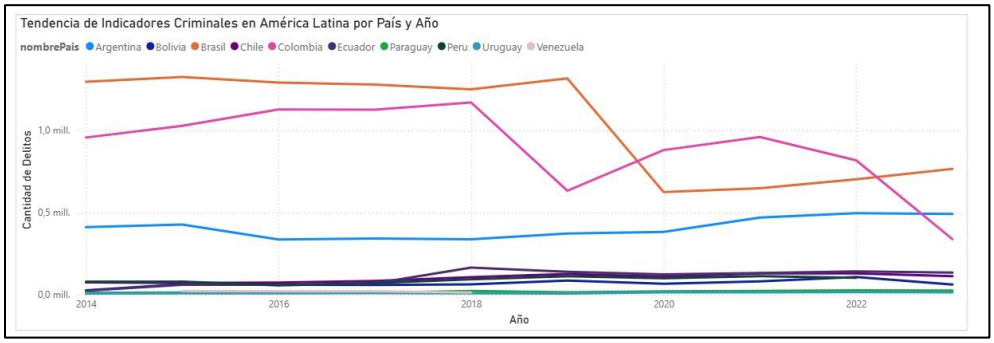
```
>
> # ------------------------------------------------------------
> # PREDICCIONES Y EVALUACIÓN
> # ------------------------------------------------------------
> predictions <- predict(svr_model, newdata = test_data)
>
> mae <- mean(abs(predictions - test_data$valorIndicador))
> mse <- mean((predictions - test_data$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
```

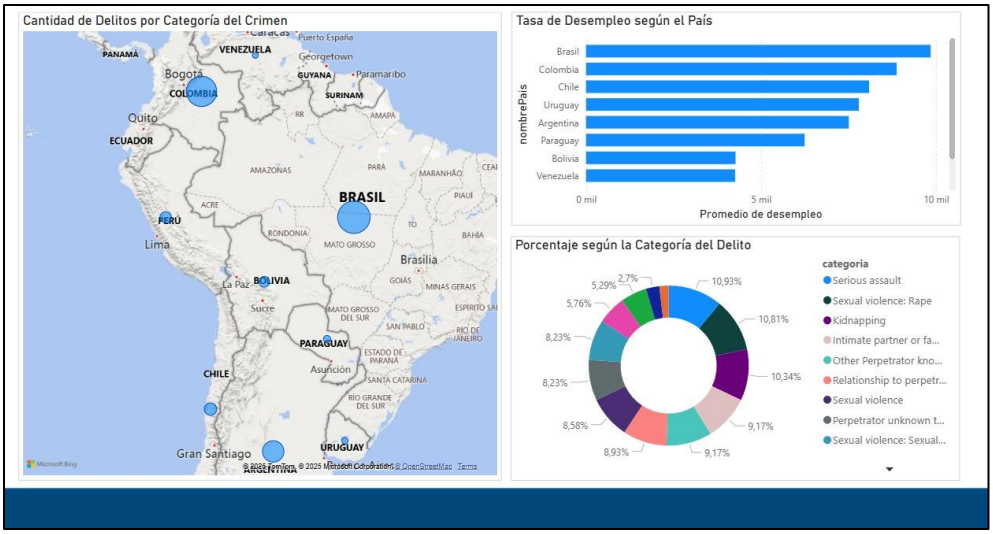


5. Visualización de Datos

Para complementar el análisis numérico, se elaboraron diferentes gráficos que facilitan la comprensión visual de la información.







6. Conclusiones

El análisis de la criminalidad en América Latina entre 2013 y 2025 confirma la existencia de diferencias significativas en la incidencia delictiva según países, años y categorías de delito, revelando un panorama complejo y heterogéneo. Se encontró que las variables socioeconómicas, incluido el desempleo y el PIB total, eran factores determinantes que influían significativamente en el número total de delitos registrados (valor del indicador), lo que respalda la hipótesis de que las condiciones económicas desfavorables pueden aumentar la delincuencia.

Respecto a los modelos predictivos, los métodos que capturan relaciones no lineales y complejas, como Random Forest y XGBoost, mostraron el mejor desempeño, con una capacidad explicativa superior al 80% de la varianza y errores mínimos (MAE menor a 0,8), demostrando alta precisión y robustez en la predicción de los niveles de criminalidad. El análisis de importancia variable utilizando Random Forest mostró que factores como el país y el salario mínimo tienen un impacto significativo, orientando las políticas públicas futuras.

Por el contrario, los modelos lineales más simples, como la regresión múltiple y los árboles de decisión, si bien son útiles para interpretar los efectos directos y las jerarquías entre

variables, tienen limitaciones para captar la complejidad del fenómeno, lo que resulta en errores de predicción mayores. La regresión SVR, aunque flexible, mostró un desempeño moderado, lo que podría explicarse por la necesidad de una mayor optimización o la naturaleza multifactorial de los datos.

En general, estos resultados indican que la incorporación de variables socioeconómicas en modelos avanzados de aprendizaje automático es una estrategia eficaz para comprender y predecir la delincuencia, y sugieren que las políticas centradas en mejorar las condiciones económicas pueden ayudar a reducir las tasas de delincuencia en la región.

7. Anexos

- Código completo utilizado para el análisis.

Adjunto en el entregable.

Bibliografía

Banco Mundial. (2023). *World Development Indicators*. Recuperado de https://databank.worldbank.org/source/world-development-indicators

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (R package version 1.7-9) [Software]. https://cran.r-project.org/package=e1071

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.1) [Software]. R Foundation for Statistical Computing. https://www.R-project.org/