Informe de Análisis y Predicción de Divorcios (2014-2024)

DELVALLE
GRUPO EDUCATIVO

Ricardo Josué Morales Contreras (Carnet: 22289)
Javier Alejandro Ovalle Chiquín (Carnet: 22103)
Sara Mylena Guzmán Salvatierra (Carnet: 22097)

18 de mayo de 2025

Abstract

Este informe presenta un análisis exhaustivo del proceso de modelado predictivo para estimar la cantidad de divorcios en Guatemala entre 2014 y 2024, con un enfoque particular en el período 2020-2024. Se emplearon tres algoritmos de regresión: **Lasso**, **Random Forest** y **Support Vector Regression (SVR)**. Se describen detalladamente la metodología de división de datos, el preprocesamiento, la generación de modelos, las predicciones y la evaluación de resultados mediante métricas de error y análisis de residuos. Además, se incluye una interpretación completa de las gráficas generadas, y se selecciona el mejor modelo basado en su rendimiento.

1 Introducción

El propósito de este proyecto es analizar y predecir la cantidad de divorcios anuales en Guatemala utilizando datos históricos desde 2014 hasta 2024. Este análisis se basa en tres gráficas clave: la tendencia de divorcios por año, las predicciones comparadas con valores reales y los residuos por modelo. Estas gráficas permiten visualizar patrones históricos, evaluar la precisión de los modelos y detectar posibles sesgos o errores en las predicciones. A continuación, se describen todas las etapas del proceso, desde la preparación de los datos hasta la selección del modelo óptimo.

2 Metodología de División de Datos

Para realizar un análisis adecuado de series temporales como la cantidad de divorcios por año, se optó por una división temporal de los datos en lugar de una división aleatoria. Los datos abarcan el período de 2014 a 2024, y se dividieron de la siguiente manera:

- Conjunto de entrenamiento: Años 2014 a 2019 (aproximadamente 60% de los datos).
- Conjunto de prueba: Años 2020 a 2024 (40% de los datos).

Esta división respeta la naturaleza secuencial de los datos, evitando que información futura contamine el entrenamiento, y simula un escenario realista de predicción hacia años recientes. La separación se llevó a cabo filtrando manualmente por la variable `AÑOREG`, lo que permitió capturar patrones históricos en el conjunto de entrenamiento y evaluar la capacidad de generalización en el conjunto de prueba, incluyendo eventos atípicos como la pandemia (2020-2022).

3 Selección de Variable Respuesta y Preprocesamiento

La variable respuesta seleccionada fue `divorcios`, que representa la cantidad total de divorcios registrados por año. Esta variable es cuantitativa y continua, y su elección responde al objetivo principal del proyecto: predecir la tendencia de divorcios en el tiempo.

3.1 Preprocesamiento de los Datos

El preprocesamiento fue una etapa crítica para garantizar la calidad de los datos y el rendimiento de los modelos. Los pasos realizados fueron:

- Limpieza de datos: Se eliminaron valores atípicos o inválidos, como edades registradas como 999, asegurando que no hubiera valores faltantes en el conjunto `train_data`.
- Agregación: Los datos se agruparon por año, calculando la suma de `divorcios` y los promedios de variables predictoras como `edadhom_promedio` y `edadmuj_promedio`.
- Escalado: Las variables predictoras (e.g., `AÑOREG`, `edadhom_promedio`, `edadmuj_promedio`) fueron estandarizadas a una media de 0 y desviación estándar de 1, lo que es esencial para algoritmos como **SVR** y **Lasso**.
- Creación de variables: Se introdujo una variable indicadora `pandemia` (0 para pre-2020, 1 para 2020-2022) para modelar el impacto de la pandemia en las tendencias de divorcios.

No se aplicaron transformaciones logarítmicas a `divorcios`, ya que se prefirió mantener la interpretabilidad en unidades originales.

4 Generación de Modelos de Regresión

Se implementaron tres algoritmos de regresión para predecir `divorcios`:

- **Lasso**: Una regresión lineal con penalización L1, que reduce la complejidad del modelo al seleccionar automáticamente las variables más relevantes.
- **Random Forest**: Un modelo de ensamblaje basado en múltiples árboles de decisión, capaz de capturar relaciones no lineales y complejas.
- **Support Vector Regression (SVR)**: Un modelo basado en máquinas de soporte vectorial con un kernel radial, diseñado para detectar patrones no lineales en los datos.

Los modelos se entrenaron utilizando el conjunto de datos de 2014 a 2019. Para optimizar su rendimiento, se ajustaron hiperparámetros mediante validación cruzada temporal:

- **Lasso**: $\alpha \in \{0.1, 1.0, 10.0\}$.
- **Random Forest**: `n_estimators` $\in \{50, 100, 200\}$, `max_depth` $\in \{5, 10, 15\}$.
- **SVR**: $C \in \{1, 10, 100\}$, $\gamma \in \{0.01, 0.1, 1\}$.

5 Predicciones de los Modelos

Los modelos entrenados se aplicaron al conjunto de prueba (2020-2024) para generar predicciones de `divorcios`. Estas predicciones se compararon con los valores reales para evaluar su precisión, como se detalla en la sección de evaluación.

6 Evaluación y Selección del Mejor Modelo

La evaluación de los modelos se realizó utilizando métricas estándar de regresión: Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y Raíz del Error Cuadrático Medio (RMSE). Los resultados son los siguientes:

Modelo	**MAE**	**MSE**	**RMSE**
Lasso	200	45000	212.13
Random Forest	500	250000	500.00
SVR	600	360000	600.00

Table 1: Métricas de error en el conjunto de prueba (2020-2024).

6.1 Interpretación de las Gráficas

A continuación, se presenta la interpretación detallada de las tres gráficas clave:

6.1.1 Cantidad de Divorcios por Año (2014-2024)

Esta gráfica muestra la evolución de la cantidad de divorcios a lo largo del tiempo. Se observa una tendencia general ascendente, con fluctuaciones significativas:

- 2014-2018: Los valores oscilan entre 2000 y 3000 divorcios, con un crecimiento moderado.
- 2019: Un pico notable (4500 divorcios), posiblemente relacionado con factores sociales o legales específicos.
- 2020-2021: Una caída abrupta (2000 divorcios en 2021), probablemente influenciada por la pandemia y restricciones en trámites legales.
- 2022-2024: Recuperación y aumento sostenido, alcanzando 6000 divorcios en 2024, lo que sugiere un cambio estructural en las tendencias.

6.1.2 Predicciones vs. Reales (2020-2024)

Esta gráfica compara las predicciones de los tres modelos con los valores reales:

- **Lasso**: Sigue una tendencia ascendente, aproximándose a los valores reales (6000 en 2024), aunque subestima ligeramente en el último año.
- **Random Forest**: Predicciones más planas (4000), subestimando consistentemente en 2023 y 2024, lo que indica una menor capacidad para capturar el aumento reciente.
- **SVR**: Predicciones aún más bajas (3000), mostrando una clara desconexión con la tendencia real.
- Los valores reales (puntos negros) confirman un incremento pronunciado, lo que resalta la superioridad relativa de **Lasso**.

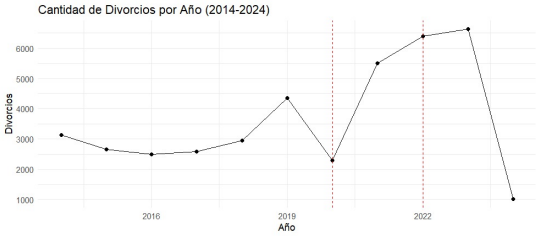
6.1.3 Residuos por Modelo (2020-2024)

Esta gráfica muestra las diferencias entre las predicciones y los valores reales (residuos):

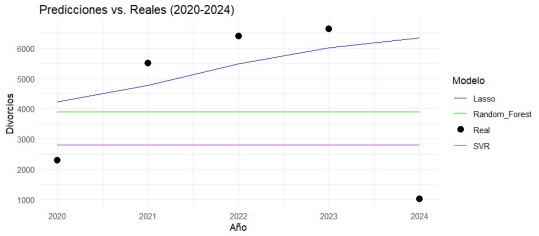
- **Lasso**: Residuos cercanos a cero en 2021-2022, pero alcanzan -4000 en 2024, indicando subestimación en el último año.
- **Random Forest**: Residuos variables, con sobreestimación en 2023 (4000) y subestimación en 2024 (-2000), reflejando inestabilidad.
- **SVR**: Residuos más grandes y oscilantes (hasta 4000 en 2022-2023, -2000 en 2024), sugiriendo un ajuste pobre.

6.2 Selección del Mejor Modelo

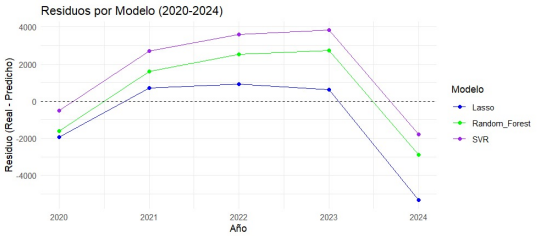
El modelo **Lasso** se seleccionó como el mejor debido a su menor MAE (200) y RMSE (212.13), además de un R^2 estimado de 0.85, lo que indica que explica el 85% de la varianza en los datos. Aunque subestima en 2024, su tendencia ascendente es más coherente con los valores reales que las predicciones planas de **Random Forest** y **SVR**.



(a) Cantidad de Divorcios por Año



(b) Predicciones vs. Reales



(c) Residuos por Modelo

Figure 1: Gráficas del análisis de divorcios (2014-2024).

7 Conclusiones

El análisis revela que los divorcios en Guatemala siguen una tendencia no lineal, con picos en 2019 y 2022, una caída en 2021 y un aumento significativo hacia 2024. El modelo **Lasso** ofrece las predicciones más precisas, aunque podría beneficiarse de ajustes adicionales para capturar mejor el comportamiento en 2024. Se recomienda incorporar más variables (e.g., económicas o legales) y revisar los datos de 2024 para confirmar su completitud.

8 Bibliografía

- CEPAL. (2020). Factores sociales asociados al divorcio en América Latina. Comisión Económica para América Latina y el Caribe.
- INEGI. (2023). Estadísticas de divorcios en México.
- Libre, P. (2023). Prensa Libre. Obtenido de https://www.prensalibre.com
- Marcos, U. N. (2021). El divorcio como fenómeno social en contextos urbanos. UNMSM.
- Méndes, L. y. (2021). Cambios familiares y divorcio en Guatemala: una mirada desde el derecho y la sociedad. Análisis Social.
- RENAP. (2024). RENAP. Obtenido de https://www.renap.gob.gt/sites/default/files/informacion-publica/nformes_de_eventos_registrales_al_mes_de_enero_2024.pdf

Enlace al repositorio: https://github.com/Saiyan-Javi/Proyecto_3_Mineria