REPORTE DE RESULTADOS DE PREDICCIÓN DE LA VARIABLE CORRUPCIÓN AMPLIA

Detalles sobre el presente reporte

- Fecha: 15 de febrero de 2024
- Nivel de observación: año inicial del reporte por municipalidad y año.
- Variables predictoras: SIAF, Renamu y variables políticas
- Variable predicha: corrupción amplia
- Periodo en el que fue entrenado el modelo: 2016-2020
- Tipo de predicción: clasificación
- Ejecución: N°10

Etapas de preprocesamiento

1. **Imputación de las variables de SIAF.** Se imputó a todas las variables provenientes de la base de datos SIAF con el valor de 0.
2. **Filtro de valores perdidos**. Se descartaron todas aquellas variables con un porcentaje de valores perdidos mayor o igual al umbral de 0.1.
3. **Imputación de variables de Renamu**. Se imputó a todas las variables provenientes de la base de datos Renamu. Las variables discretas fueron imputadas con moda, y las variables continuas, con media.
4. **Filtro de variabilidad**. Se descartaron todas aquellas variables constantes, es decir, con una variabilidad de 0.
5. **Imputación de outliers**. En este paso se considera solamente las variables de SIAF. Se imputaron los valores superiores al percentil 99% con el valor del percentil 99%.
6. **Transformaciones logarítmicas**. En este paso se considera solamente las variables de SIAF y consta de 3 etapas. Primero, se identifica todas las variables con al menos un valor negativo, y se las divide entre 1 millón. Segundo, se suma 1 a todas las variables de SIAF para evitar que los valores a ser transformados logarítmicamente tomen valores negativos. Tercero, se aplica la transformación logarítmica

Número de variables

La tabla 1 presenta información sobre el número de variables en la base de datos empleada, cuyo nivel de observación es año inicial del reporte por municipalidad y año.

Tabla 1. Número de variables antes y después del preprocesamiento

Fuente	Número de variables antes del preprocesamiento	Número de variables después del preprocesamiento
SIAF	17 550	14 300
RENAMU	737	217
Variables políticas	4	4
Total	18 291	14 521
Fuente: elaboración propia		

Métodos de (re)muestreo

Se implementaron tres métodos de (re)muestreo sobre el conjunto de entrenamiento para balancear el número de observaciones por categoría de predicción. El conjunto de prueba mantiene su proporción original.

Tabla 2. Número de observaciones por categoría de predicción según método de (re)muestreo

Método de muestreo	Total de observaciones	Total de observaciones en las que sí ocurrió corrupción amplia	Total de observaciones en las que no ocurrió corrupción amplia
Original	967	896	71
SMOTE	1 792	896	896
SMOTE Tomek-Links	1 686	843	843
Naive Random Oversampling	1 792	896	896
Fuente: elaboración propia			

Hiperparámetros considerados en el Grid Search

Se utilizó el algoritmo gridsearchcv para realización una búsqueda exhaustiva de la mejor combinación de hiperparámetros (Grid Search). Los rangos de hiperparámetros considerados se presentan en las tablas 3 y 4.

Tabla 3. Hiperparámetros considerados en el Grid Search de los Métodos Basados en Árboles

Modelo	n_estimators	max_depth	max_features
Random Forest	250, 500 y 1000	20 y 30	20%, 30%, 40%
Gradient Boosting Trees	250, 500 y 1000	1 y 2	20%, 30%, 40%
LGBM Classifier	250, 500 y 1000	1 y 2	-
Regression Forest	252, 500 y 1000	10, 20 y 30	-
Fuente: elaboración propia			

Tabla 4. Hiperparámetros considerados en el Grid Search de los Métodos de Regularización

Modelo	Cs (Fuerza de la regularización)
Lasso	De 10^8 a 10^{-6}, 100 valores en escala logarítmica
Ridge	De 10^8 a 10^{-6}, 100 valores en escala logarítmica
Elastic Net	De 10^8 a 10^{-6}, 100 valores en escala logarítmica
Fuente: elaboración propia	

También debe considerarse que en el Grid Search se empleó, para todos los modelos, una validación cruzada en K-Folds, donde k siempre tuvo el valor de 5. La métrica de desempeño usada para comparar los distintos modelos durante el Grid Search fue F1 (a excepción del método Regression Forest, donde se usó el R2).

Resultados (métricas de desempeño)

La tabla 5 presenta los resultados de los modelos de Machine Learning para el conjunto de entrenamiento NRO. Tomando en cuenta la métrica F1, el modelo con el mejor desempeño es el modelo **Regression Forest** entrenado con el conjunto de entrenamiento Naive Random Oversampling (NRO). Las combinaciones óptimas de hiperparámetros se reportan en los anexos 1 y 2.

Tabla 5. Métricas de desempeño de los modelos entrenados con el conjunto de entrenamiento Naive Random Oversampling.

Métrica	Regresión Logística	Lasso	Ridge	Elastic Net	Random Forest	Gradient Boosting Trees	LGBM Classifier	Regression Forest (threshold = 0.5)
F1	0.415	0.388	0.388	0.388	0.593	0.500	0.475	0.648
Accuracy	0.530	0.458	0.458	0.458	0.906	0.906	0.906	0.901
AUC ROC	0.565	0.649	0.649	0.649	0.726	0.750	0.665	0.808
F1 (Sí)	0.674	0.595	0.595	0.595	0.950	0.951	0.951	0.947
F1 (No)	0.156	0.182	0.182	0.182	0.235	0.049	0	0.349
Fuente: elaboración propia								

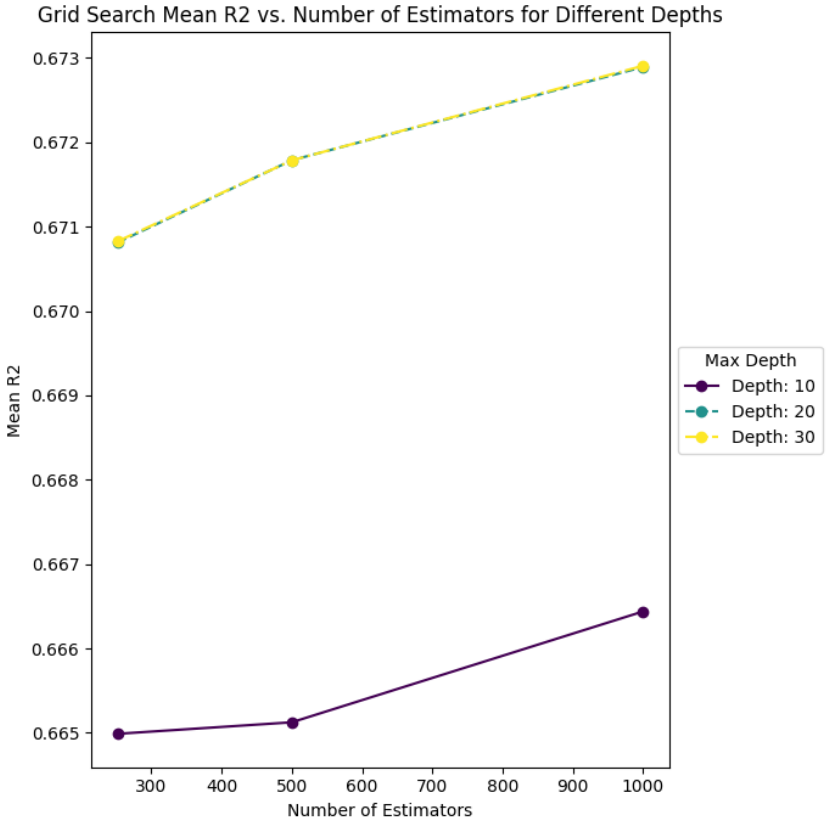
Asimismo, se presentan los 5 modelos con mejor considerando los distintos métodos de remuestreo empleados (SMOTE, SMOTE Tomek-Links y Naive Random Oversampling). Este ranking se realiza teniendo en cuenta la métrica F1:

Tabla 6. Cinco modelos con mayor poder predictivo considerando los distintos métodos de remuestreo

Modelo	F1
Regression Forest NRO (threshold = 0.5)	0.648
Regression Forest SMOTE Tomek-Links (threshold = 0.5)	0.647
Regression Forest SMOTE (threshold = 0.6)	0.635
Regression Forest NRO (threshold = 0.6)	0.634
Regression Forest O (threshold = 0.9)	0.621
Fuente: elaboración propia	

El gráfico 1. Muestra el ajuste del modelo óptimo (**Regression Forest** entrenado con el conjunto de Naive Random Oversampling) a través de los distintos hiperparámetros considerados durante el proceso de Grid Search.

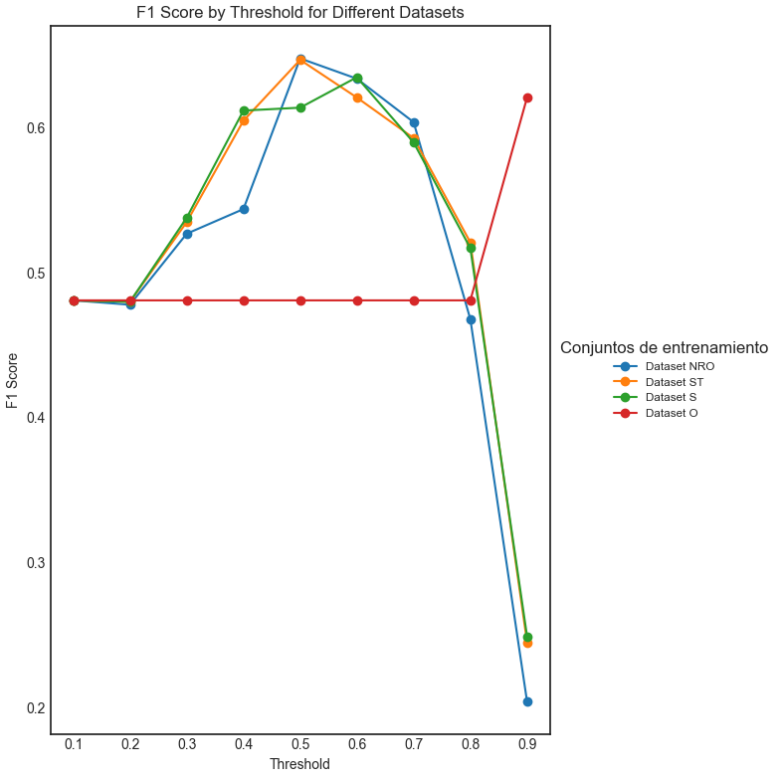
Gráfico 1. Grid Search R2 vs. Grid Search Parameters for the NRO Regression Forest Model



Modelo Regression Forest a distintos thresholds

Aunque está diseñado para tareas de regresión, el modelo Regression Forest genera predicciones que varían de 0 a 1. Las métricas anteriormente reportadas aplican un threshold de 0.5 para categorizar las predicciones en dos clases y calcular las métricas. A continuación se presenta un gráfico de cómo varía la métrica F1 en función de distintos thresholds en el rango [0.1, 0.9].

Gráfico 2. F1 en función de distintos thesholds para distintos conjuntos de entrenamiento.



Variables más importantes según el criterio de impureza de Gini

En esta sección se presentan las 20 variables más importantes según el criterio de impureza de Gini (estimado mediante el comando feature importance) para el modelo óptimo.

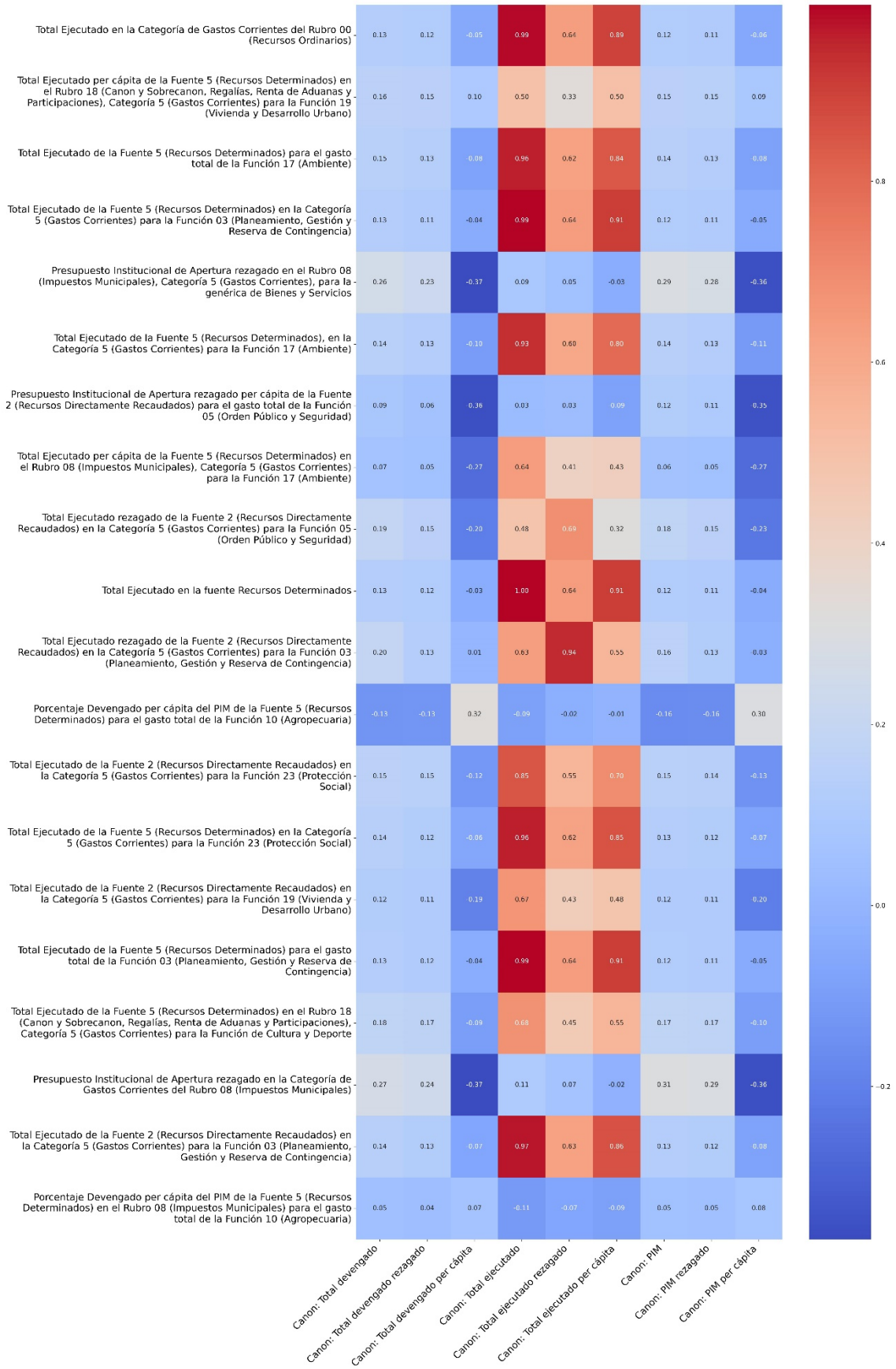
Tabla 7. Veinte variables más importantes de acuerdo con el criterio de impureza de Gini

Variable	Etiqueta	Fuente	Importance Score
tejgct_r00gstcr	Total Ejecutado en la Categoría de Gastos Corrientes del Rubro 00 (Recursos Ordinarios)	SIAF	0.082
tejgfun_f5r18ct05vivpc	Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones), Categoría 5 (Gastos Corrientes) para la Función 19 (Vivienda y Desarrollo Urbano)	SIAF	0.063
tejgtotfun_f5amb	Total Ejecutado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 17 (Ambiente)	SIAF	0.054
tejgfun_f5ct05pgrco	Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.053
_piagge_r08ct05biser	Presupuesto Institucional de Apertura rezagado en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios	SIAF	0.035
tejgfun_f5ct05amb	Total Ejecutado de la Fuente 5 (Recursos Determinados), en la Categoría 5 (Gastos Corrientes) para la Función 17 (Ambiente)	SIAF	0.034
_piagtotfun_f2opsegpc	Presupuesto Institucional de Apertura rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función 05 (Orden Público y Seguridad)	SIAF	0.022
tejgfun_f5r08ct05ambpc	Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes) para la Función 17 (Ambiente)	SIAF	0.022
_tejgfun_f2ct05opseg	Total Ejecutado rezagado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad)	SIAF	0.022

tejgft_rdet	Total Ejecutado en la fuente Recursos Determinados	SIAF	0.018
_tejgfun_f2ct05pgrco	Total Ejecutado rezagado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.017
devppimtotfun_f5agropc	Porcentaje Devengado per cápita del PIM de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 10 (Agropecuaria)	SIAF	0.017
tejgfun_f2ct05prots	Total Ejecutado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social)	SIAF	0.016
tejgfun_f5ct05prots	Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social)	SIAF	0.014
tejgfun_f2ct05viv	Total Ejecutado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 19 (Vivienda y Desarrollo Urbano)	SIAF	0.012
tejgtotfun_f5pgrco	Total Ejecutado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.012
tejgfun_f5r18ct05cydep	Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones), Categoría 5 (Gastos Corrientes) para la Función de Cultura y Deporte	SIAF	0.011
_piagct_r08gstcr	Presupuesto Institucional de Apertura rezagado en la Categoría de Gastos Corrientes del Rubro 08 (Impuestos Municipales).	SIAF	0.008
tejgfun_f2ct05pgrco	Total Ejecutado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 03 (Planeamiento, Gestión y Reserva de Contingencia).	SIAF	0.008
devppimtotfun_f5r08agropc	Porcentaje Devengado per cápita del PIM de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para el gasto total de la Función 10 (Agropecuaria).	SIAF	0.008

Asimismo, se visualiza la correlación entre las 20 variables más importantes según el criterio de impureza de Gini y las variables de Canon.

Gráfico 3. Correlación entre las 20 variables más importantes según el criterio de impureza de Gini y las variables de Canon.



Variables más importantes según el criterio de SHAP Values

En esta sección se presentan las 20 variables más importantes según el criterio de SHAP Values para el modelo óptimo.

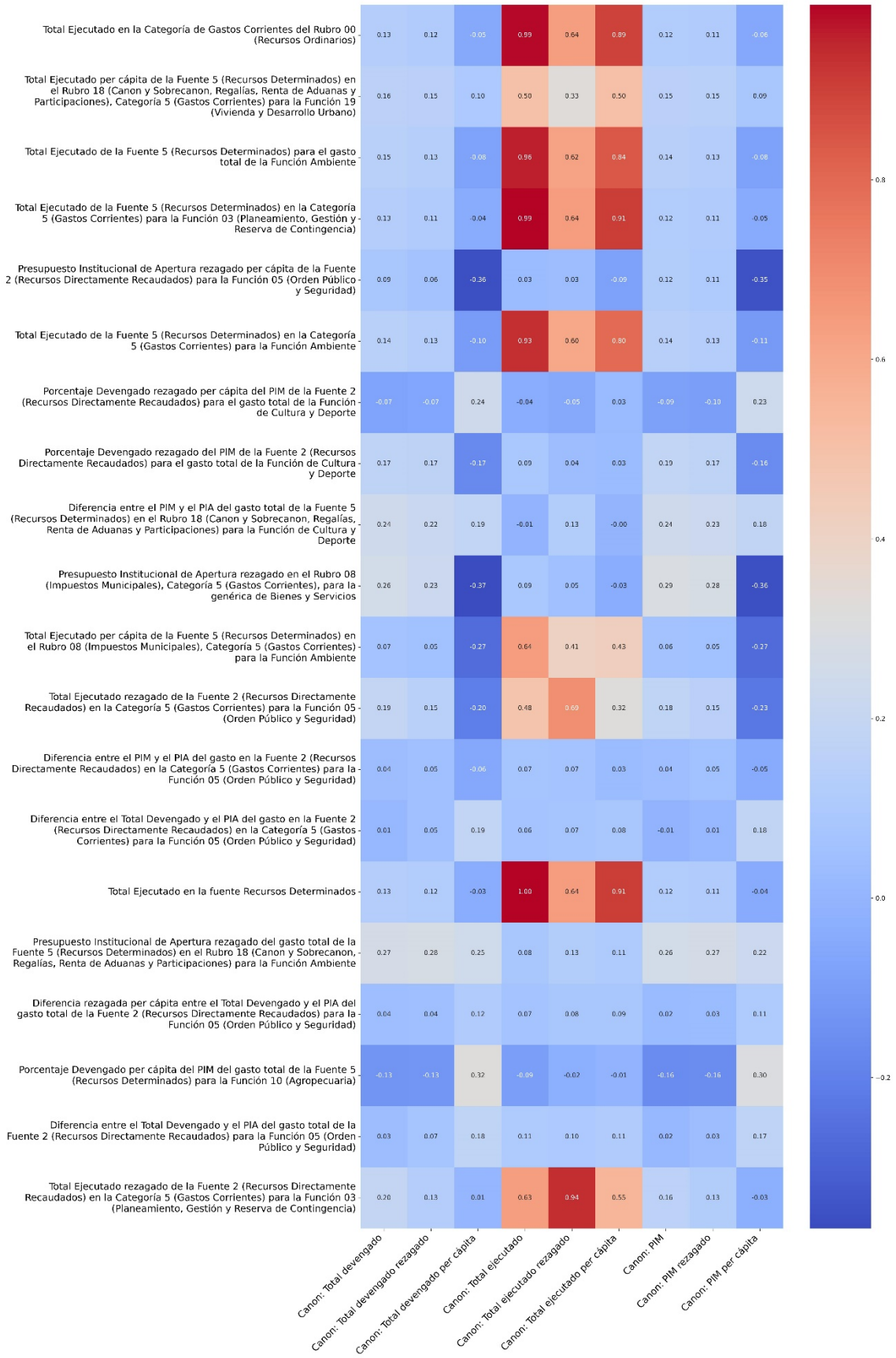
Tabla 8. Veinte variables más importantes de acuerdo con el criterio de SHAP Values.

Variable	Etiqueta	Fuente	SHAP Values
tejgct_r00gstcr	Total Ejecutado en la Categoría de Gastos Corrientes del Rubro 00 (Recursos Ordinarios)	SIAF	0.02
tejgfun_f5r18ct05vivpc	Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones), Categoría 5 (Gastos Corrientes) para la Función 19 (Vivienda y Desarrollo Urbano)	SIAF	0.016
tejgtotfun_f5amb	Total Ejecutado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función Ambiente	SIAF	0.013
tejgfun_f5ct05pgrco	Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.012
_piagtotfun_f2opsegpc	Presupuesto Institucional de Apertura rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) para la Función 05 (Orden Público y Seguridad)	SIAF	0.008
tejgfun_f5ct05amb	Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función Ambiente	SIAF	0.008
_devppimtotfun_f2cydeppc	Porcentaje Devengado rezagado per cápita del PIM de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función de Cultura y Deporte	SIAF	0.007
_devppimtotfun_f2cydep	Porcentaje Devengado rezagado del PIM de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función de Cultura y Deporte	SIAF	0.007
dfgpimpiatotfun_f5r18cydep	Diferencia entre el PIM y el PIA del gasto total de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función de Cultura y Deporte	SIAF	0.006
_piagge_r08ct05biser	Presupuesto Institucional de Apertura rezagado en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios	SIAF	0.005
tejgfun_f5r08ct05ambpc	Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes) para la Función Ambiente	SIAF	0.005
_tejgfun_f2ct05opseg	Total Ejecutado rezagado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad)	SIAF	0.005
dfgpimpiafun_f2ct05opseg	Diferencia entre el PIM y el PIA del gasto en la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad)	SIAF	0.005
dfgdevpiagfun_f2ct05opseg	Diferencia entre el Total Devengado y el PIA del gasto en la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad)	SIAF	0.005
tejgft_rdet	Total Ejecutado en la fuente Recursos Determinados	SIAF	0.005
_piagtotfun_f5r18amb	Presupuesto Institucional de Apertura rezagado del gasto total de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Ambiente	SIAF	0.004
_dfgdevpiagtotfun_f2opsegpc	Diferencia rezagada per cápita entre el Total Devengado y el PIA del gasto total de la Fuente 2 (Recursos Directamente Recaudados) para la Función 05 (Orden Público y Seguridad)	SIAF	0.004
devppimtotfun_f5agropc	Porcentaje Devengado per cápita del PIM del gasto total de la Fuente 5 (Recursos Determinados) para la Función 10 (Agropecuaria)	SIAF	0.004
dfgdevpiagtotfun_f2opseg	Diferencia entre el Total Devengado y el PIA del gasto total de la Fuente 2 (Recursos Directamente Recaudados) para la Función 05 (Orden Público y Seguridad)	SIAF	0.004

_tejgfun_f2ct05pgrco	Total Ejecutado rezagado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.004

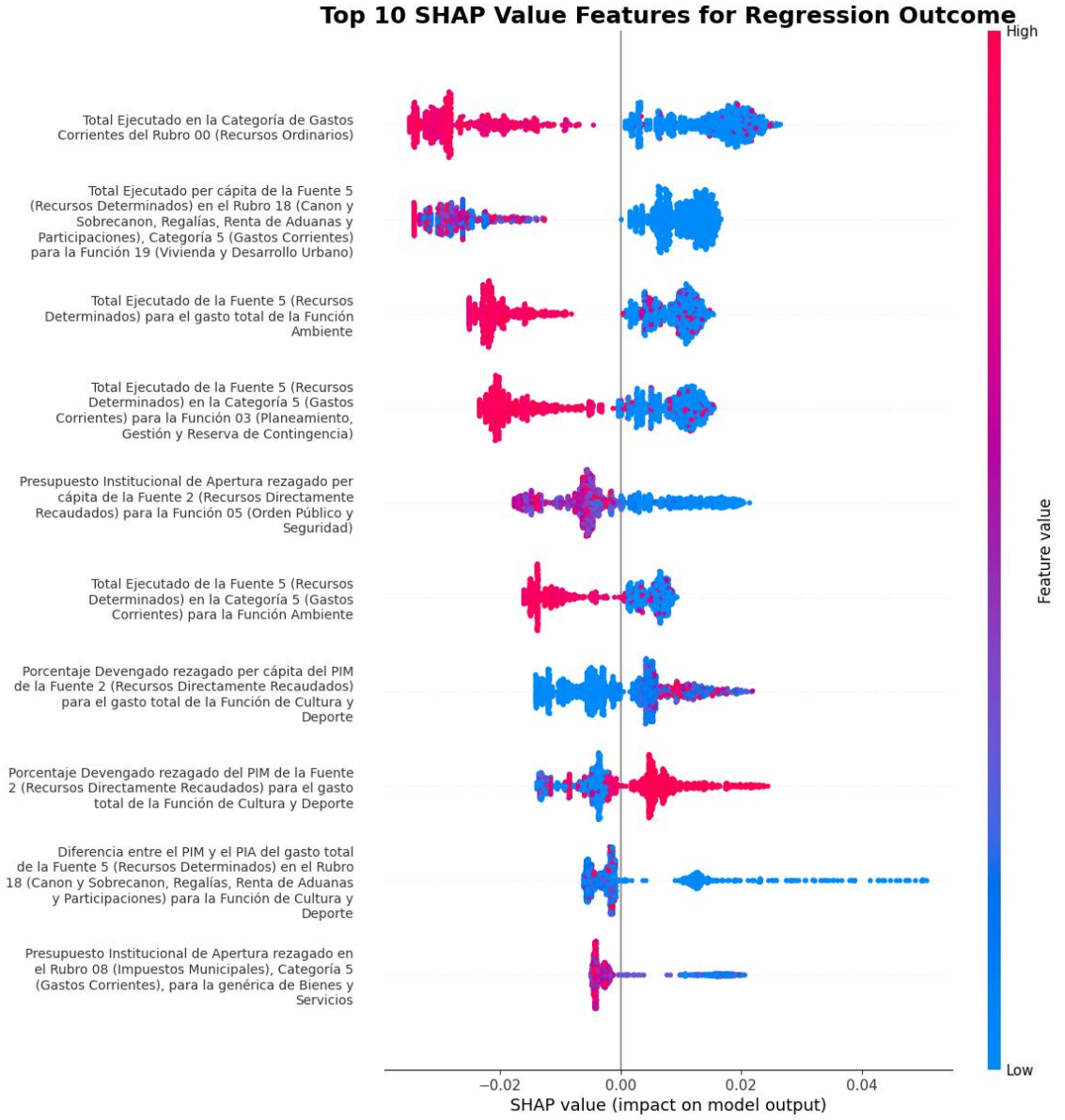
Asimismo, se visualiza la correlación entre las 20 variables más importantes según el criterio SHAP Values y las variables de Canon.

Gráfico 4. Correlación entre las 20 variables más importantes según el criterio de SHAP Values y las variables de Canon.



Adicionalmente se presenta un gráfico de SHAP Values que indica cómo cada una de las 10 variables más importantes influye en la predicción de casos de corrupción amplia para el modelo óptimo seleccionado. Cada punto representa una observación: los puntos rojos indican valores altos de la característica y los azules, valores bajos. La posición horizontal de los puntos refleja la magnitud de la influencia de la característica en la predicción. Las características están ordenadas de mayor a menor impacto en el eje vertical. Un punto hacia la derecha sugiere que la característica incrementa la probabilidad de corrupción amplia, mientras que un punto hacia la izquierda sugiere lo contrario. La concentración de puntos muestra la variabilidad de la influencia de la característica: una mayor dispersión indica mayor variabilidad en su impacto en las predicciones.

Gráfico 5. 10 variables más importantes según criterio SHAP para casos positivos de Corrupción Amplia



Reentranamiento del modelo óptimo

Una vez seleccionado el modelo óptimo (Regression Forest NRO), se seleccionaron las variables más importantes según el criterio de impureza de Gini a un 80% de suma acumulada. En total fueron 107 variables. Una vez seleccionadas las variables, se reentró el modelo Regression Forest NRO con los mismos hiperparámetros que en el modelo óptimo seleccionado (véase el Anexo 1). La tabla 9 presenta las métricas de desempeño sobre el conjunto de prueba. El threshold para categorizar las predicciones continuas del modelo Regression Forest es de 0.5.

Tabla 9. Métricas de desempeño para el modelo óptimo entrenado con variables seleccionadas mediante el 80% de suma acumulada según el criterio de impureza de Gini

Modelo	F1	Accuracy	ROC AUC	F1 (Sí)	F1 (No)
Regression Forest NRO (variables = 80% de suma acumulada según el criterio de impureza de Gini).	0.676	0.901	0.797	0.946	0.406
Fuente: elaboración propia					

ANEXOS

Anexo 1. Combinación óptima de hiperparámetros para modelos Naive Random Oversampling en árboles

Modelo	n_estimators	max_depth	max_features
Regression Forest	1000	30	-
LGBM Classifier	1000	2	-
Gradient Boosting Trees	1000	2	20%
Random Forest	500	20	20%
Fuente: elaboración propia			

Anexo 2. Combinación óptima de hiperparámetros para modelos Naive Random Oversampling de regularización

Modelo	Cs (Fuerza de la regularización)
Lasso	114.9757
Ridge	114.9757
Elastic Net	114.9757
Fuente: elaboración propia	