REPORTE DE RESULTADOS DE PREDICCIÓN DE LA VARIABLE CORRUPCIÓN INTENSA

1. 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

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 **Random Forest** entrenado con el conjunto de entrenamiento SMOTE. 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 SMOTE.

Métrica	Regresión Logística	Lasso	Ridge	Elastic Net	Random Forest	Gradient Boosting Trees	LGBM Classifier	Regression Forest (threshold = 0.5)
F1	0.326	0.426	0.426	0.426	0.627	0.594	0.557	0.604
Accuracy	0.340	0.427	0.427	0.427	0.725	0.723	0.701	0.643
AUC ROC	0.494	0.515	0.515	0.515	0.680	0.645	0.626	0.672
F1 (Sí)	0.230	0.447	0.447	0.447	0.818	0.823	0.810	0.729
F1 (No)	0.422	0.405	0.405	0.405	0.436	0.365	0.303	0.479
Fuente: elaboración propia								

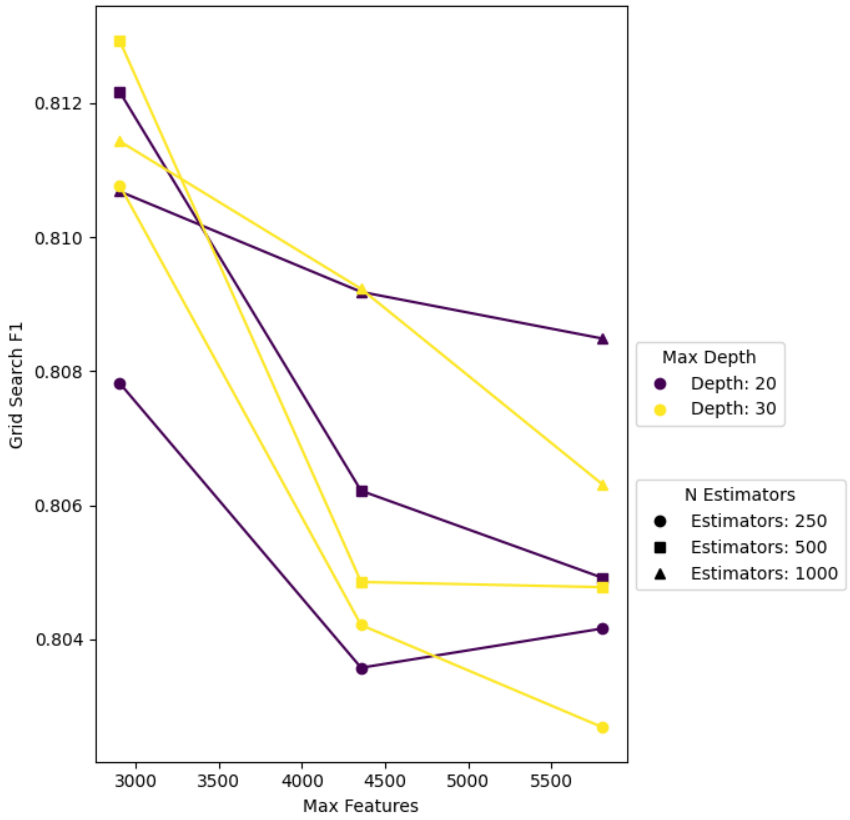
Asimismo, se presentan los 5 modelos con mejor considerando los distintos métodos de remuestreo empleados (SMOTE, SMOTE Tomek-Links y Naive Random Oversampling). Este ranking se realiza teniendo en cuenta la métrica F1:

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

Modelo	F1
Regression Forest NRO (threshold = 0.4)	0.631
Random Forest SMOTE	0.627
Regression Forest Original (threshold = 0.7)	0.626
Random Forest SMOTE Tomek-Links	0.609
Regression Forest NRO (threshold = 0.5)	0.607
Fuente: elaboración propia	

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

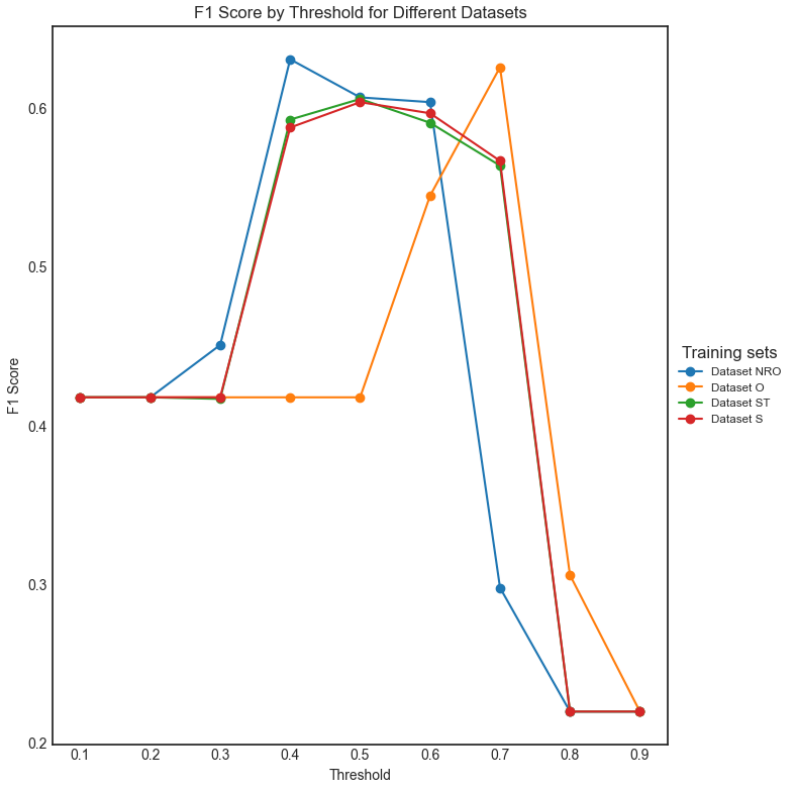
Gráfico 1. Grid Search F1 vs. Grid Search Parameters for the Random Forest SMOTE model.



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.

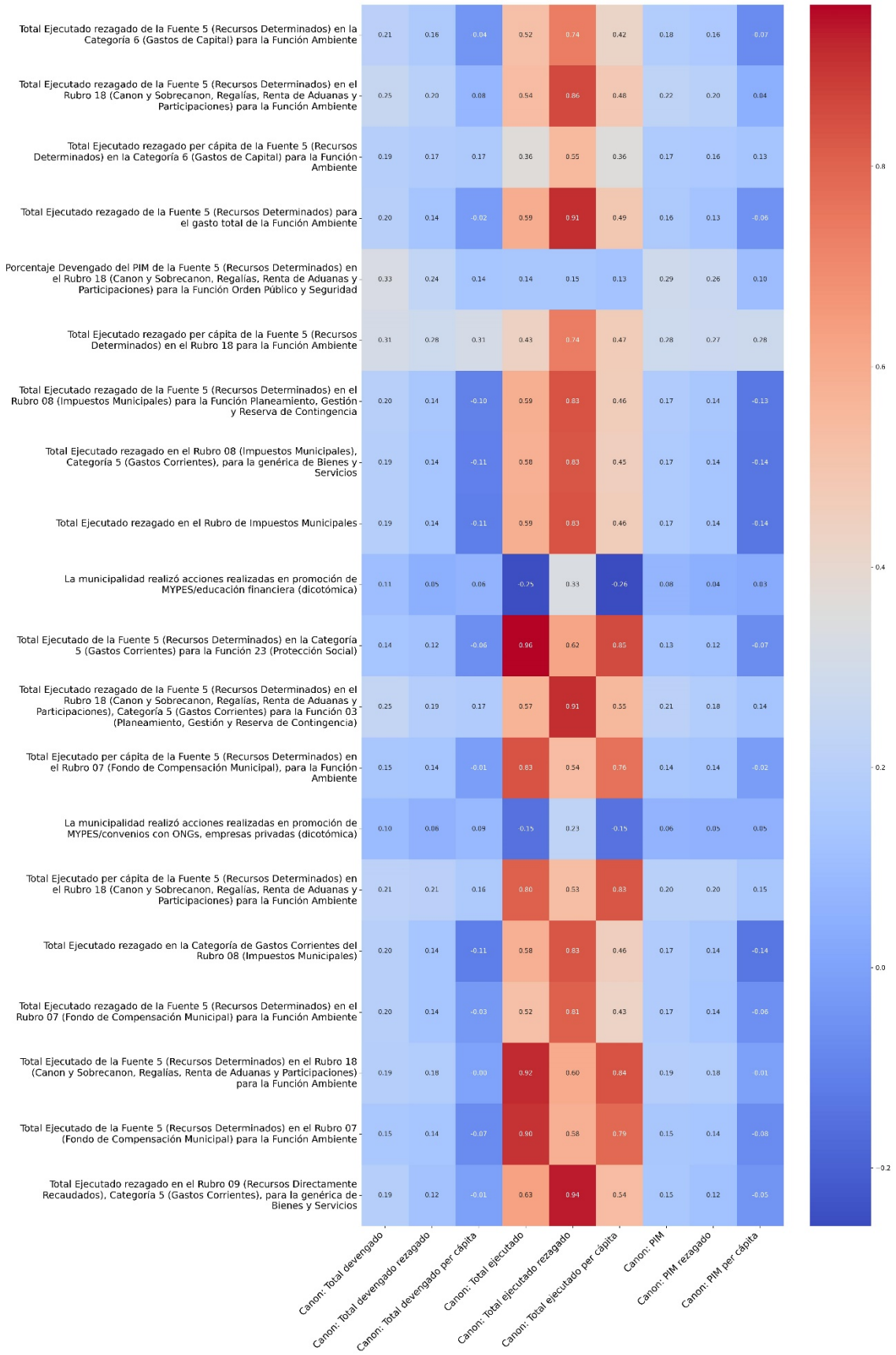
Tabla 7. Veinte variables más importantes de acuerdo con el criterio de impureza de Gini

Variable	Etiqueta	Fuente	Importance Score
_tejgfun_f5ct06amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Ambiente	SIAF	0.032
_tejgtotfun_f5r18amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Ambiente	SIAF	0.024
_tejgfun_f5ct06ambpc	Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Ambiente	SIAF	0.018
_tejgtotfun_f5amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función Ambiente	SIAF	0.011
devppimtotfun_f5r18opseg	Porcentaje Devengado del PIM de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Orden Público y Seguridad	SIAF	0.009
_tejgtotfun_f5r18ambpc	Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 para la Función Ambiente	SIAF	0.008
_tejgtotfun_f5r08pgrco	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para la Función Planeamiento, Gestión y Reserva de Contingencia	SIAF	0.007
_tejgge_r08ct05biser	Total Ejecutado rezagado en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios	SIAF	0.006
_tejgrb_impm	Total Ejecutado rezagado en el Rubro de Impuestos Municipales	SIAF	0.006

empinc_7	La municipalidad realizó acciones realizadas en promoción de MYPES/educación financiera (dicotómica)	Renamu	0.005
tejgfun_f5ct05prots	Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social)	SIAF	0.005
_tejgfun_f5r18ct05pgrco	Total Ejecutado rezagado 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 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.005
tejgtotfun_f5r07ambpc	Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal), para la Función Ambiente	SIAF	0.005
empinc_4	La municipalidad realizó acciones realizadas en promoción de MYPES/convenios con ONGs, empresas privadas (dicotómica)	Renamu	0.004
tejgtotfun_f5r18ambpc	Total Ejecutado per cápita 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
_tejgct_r08gstcr	Total Ejecutado rezagado en la Categoría de Gastos Corrientes del Rubro 08 (Impuestos Municipales)	SIAF	0.004
_tejgtotfun_f5r07amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para la Función Ambiente	SIAF	0.004
tejgtotfun_f5r18amb	Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Ambiente	SIAF	0.003
tejgtotfun_f5r07amb	Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para la Función Ambiente	SIAF	0.003
_tejgge_r09ct05biser	Total Ejecutado rezagado en el Rubro 09 (Recursos Directamente Recaudados), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios	SIAF	0.003

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

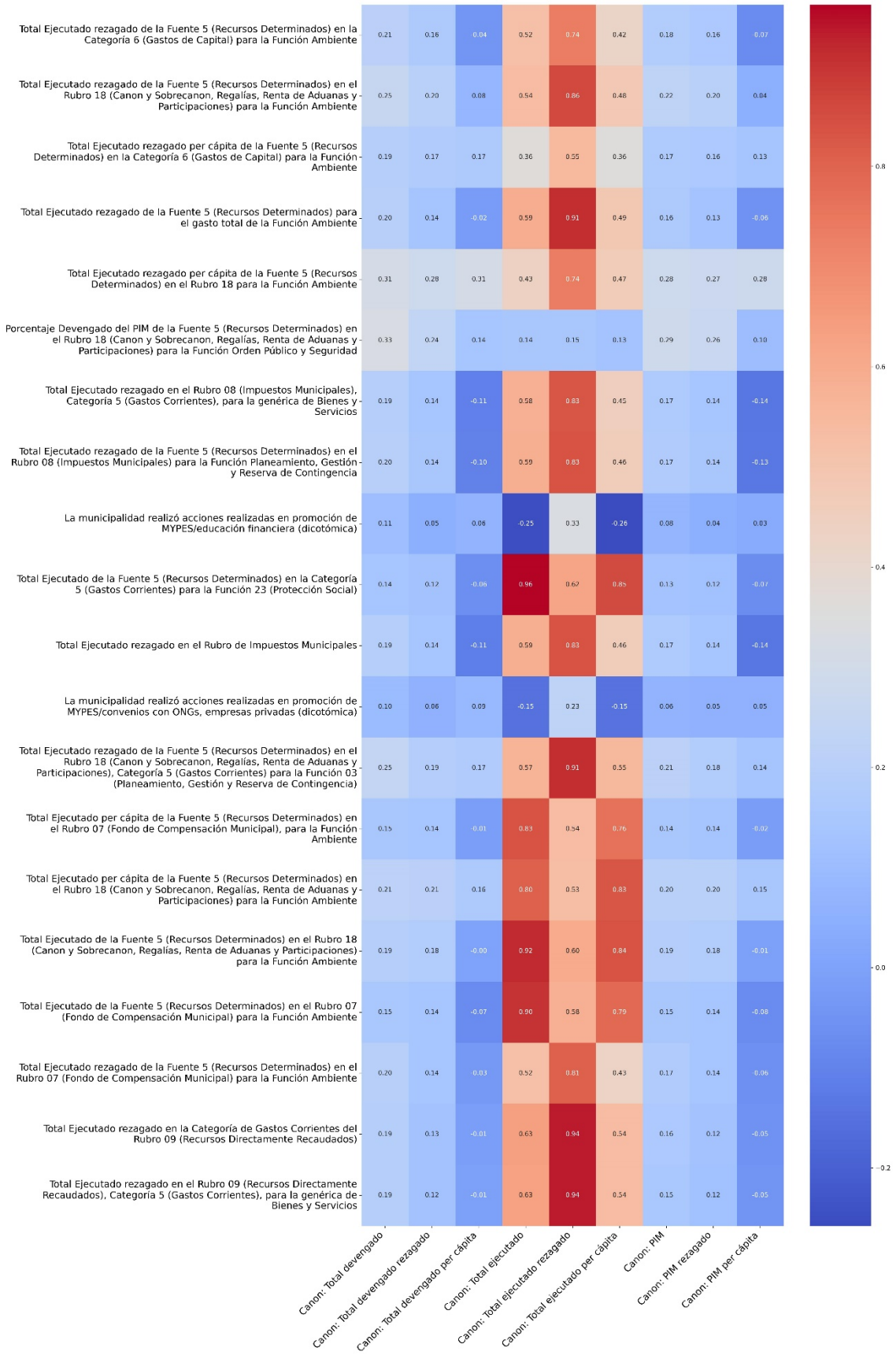
En esta sección se presentan las 20 variables más importantes según el criterio de SHAP Values para el modelo óptimo.

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

Variable	Etiqueta	Fuente	SHAP Values
_tejgfun_f5ct06amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Ambiente	SIAF	0.031
_tejgtotfun_f5r18amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Ambiente	SIAF	0.024
_tejgfun_f5ct06ambpc	Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Ambiente	SIAF	0.017
_tejgtotfun_f5amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función Ambiente	SIAF	0.012
_tejgtotfun_f5r18ambpc	Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 para la Función Ambiente	SIAF	0.008
devppimtotfun_f5r18opseg	Porcentaje Devengado del PIM de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Orden Público y Seguridad	SIAF	0.007
_tejgge_r08ct05biser	Total Ejecutado rezagado en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios	SIAF	0.007
_tejgtotfun_f5r08pgrco	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para la Función Planeamiento, Gestión y Reserva de Contingencia	SIAF	0.007
empinc_7	La municipalidad realizó acciones realizadas en promoción de MYPES/educación financiera (dicotómica)	Renamu	0.006
tejgfun_f5ct05prots	Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social)	SIAF	0.005
_tejgrb_impm	Total Ejecutado rezagado en el Rubro de Impuestos Municipales	SIAF	0.005
empinc_4	La municipalidad realizó acciones realizadas en promoción de MYPES/convenios con ONGs, empresas privadas (dicotómica)	Renamu	0.005
_tejgfun_f5r18ct05pgrco	Total Ejecutado rezagado 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 03 (Planeamiento, Gestión y Reserva de Contingencia)	SIAF	0.005
tejgtotfun_f5r07ambpc	Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal), para la Función Ambiente	SIAF	0.005
tejgtotfun_f5r18ambpc	Total Ejecutado per cápita 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
tejgtotfun_f5r18amb	Total Ejecutado 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
tejgtotfun_f5r07amb	Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para la Función Ambiente	SIAF	0.004
_tejgtotfun_f5r07amb	Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para la Función Ambiente	SIAF	0.003
_tejgct_r09gstcr	Total Ejecutado rezagado en la Categoría de Gastos Corrientes del Rubro 09 (Recursos Directamente Recaudados)	SIAF	0.003
_tejgge_r09ct05biser	Total Ejecutado rezagado en el Rubro 09 (Recursos Directamente Recaudados), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios	SIAF	0.003

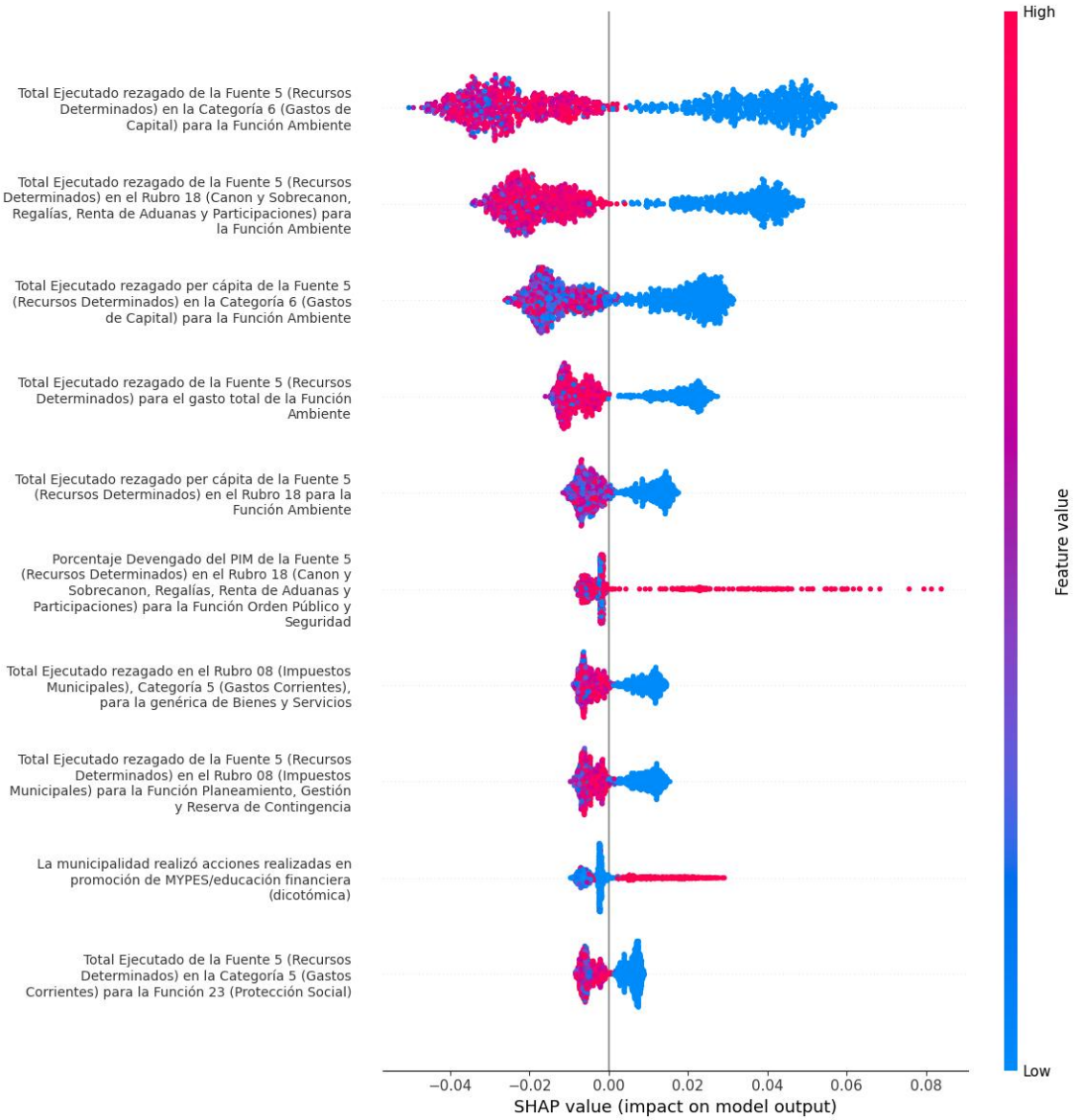
Asimismo, se visualiza la correlación entre las 20 variables más importantes según el criterio SHAP Values y las variables de Canon.

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 intensa 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 intensa, 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 Intensa



Reentranamiento del modelo óptimo

Una vez seleccionado el modelo óptimo (Random Forest SMOTE), se seleccionaron las variables más importantes según el criterio de impureza de Gini a un 80% de suma acumulada. En total fueron **3038 variables**. Una vez seleccionadas las variables, se reentró el modelo Random Forest SMOTE 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.

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)
Random Forest SMOTE (variables = 80% de suma acumulada según el criterio de impureza de Gini).	0.604	0.713	0.665	0.812	0.396
Fuente: elaboración propia					

ANEXOS

Anexo 1. Combinación óptima de hiperparámetros para modelos SMOTE en árboles

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

Anexo 2. Combinación óptima de hiperparámetros para modelos SMOTE de regularización

Modelo	Cs (Fuerza de la regularización)
Lasso	1450829
Ridge	1450829
Elastic Net	1450829
Fuente: elaboración propia	