**REPORTE DE RESULTADOS DE PREDICCIÓN DE LA VARIABLE CORRUPCIÓN AMPLIA**

**Detalles sobre el presente reporte**

* Fecha: 05 de febrero de 2024
* Nivel de observación: año inicial del reporte por municipalidad y año.
* Variables predictoras: únicamente provenientes de la base de datos de SIAF
* Variable predicha: corrupción amplia
* Periodo en el que fue entrenado el modelo: 2016-2020
* Periodo de predicción: 2009-2020
* Tipo de predicción: clasificación
* Ejecución N°8

**Número de variables**

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

**Tabla 1.** Número de variables predictoras antes y después del preprocesamiento

|  |  |  |
| --- | --- | --- |
| **Fuente** | **Número de variables predictoras antes del preprocesamiento** | **Número de variables predictoras después del preprocesamiento** |
| SIAF | 17 549 | 14 317 |
| 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 | 964 | 891 | 73 |
| SMOTE | 1 782 | 891 | 891 |
| SMOTE Tomek-Links | 1 776 | 888 | 888 |
| Naive Random Oversampling | 1 782 | 891 | 891 |
| 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 | 10, 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 | 0.001, 0.01, 0.1, 1, 10 y 100 |
| Ridge | 0.001, 0.01, 0.1, 1, 10 y 100 |
| Elastic Net | 0.001, 0.01, 0.1, 1, 10 y 100 |
| Fuente: elaboración propia | |

También debe considerarse que en el Grid Search se empleó, para todos los modelos, una validación cruzada en K-Folds, donde siempre tuvo el valor de 5.

**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** |
| F1 | 0.536 | 0.542 | 0.542 | 0.542 | 0.590 | 0.565 | 0.549 | 0.644 |
| Accuracy | 0.845 | 0.783 | 0.783 | 0.783 | 0.903 | 0.911 | 0.913 | 0.891 |
| AUC ROC | 0.566 | 0.632 | 0.632 | 0.632 | 0.734 | 0.775 | 0.684 | 0.804 |
| F1 (Sí) | 0.915 | 0.874 | 0.874 | 0.874 | 0.948 | 0.953 | 0.954 | 0.941 |
| F1 (No) | 0.158 | 0.211 | 0.211 | 0.211 | 0.231 | 0.178 | 0.143 | 0.348 |
| 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 | 0.644 |
| Regression Forest S | 0.628 |
| Regression Forest ST | 0.620 |
| Random Forest NRO | 0.590 |
| Random Forest ST | 0.569 |
| 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, Gráfico de líneas

Descripción generada automáticamente

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

**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** | **Importance Score** |
| \_devppimtotfun\_f2cydeppc | Porcentaje Devengado rezagado per cápita del PIM de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función 21, que corresponde a Cultura y Deporte | 0.083 |
| tejgtotfun\_f5viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 19 (Vivienda y Desarrollo Urbano) | 0.063 |
| 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) | 0.055 |
| tejgfun\_f5ct05viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 19 (Vivienda y Desarrollo Urbano) | 0.049 |
| tejgfun\_f5r08ct05prots | Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social) | 0.036 |
| 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) | 0.036 |
| \_tejgtotfun\_f5viv | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 19 (Vivienda y Desarrollo Urbano) | 0.034 |
| \_tejgfun\_f2ct05opsegpc | Total Ejecutado rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad) | 0.027 |
| \_piagfun\_f2ct05opsegpc | Presupuesto Institucional de Apertura rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad) | 0.027 |
| tejgge\_r09ct05otgst | Total Ejecutado en el Rubro 09 (Recursos Directamente Recaudados), Categoría 5 (Gastos Corrientes), para la genérica de Otros Gastos | 0.025 |
| tejgtotfun\_f5r08prots | otal Ejecutado de la Fuente 5 (Recursos Determinados), Rubro 08 (Impuestos Municipales) para la Función 23 (Protección Social) | 0.015 |
| tejgrb\_impm | Total Ejecutado en el Rubro de Impuestos Municipales | 0.014 |
| tejgfun\_f5r08ct05protspc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados), en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social) | 0.013 |
| \_tejgge\_r08ct05popso | Total Ejecutado rezagado en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes), para la genérica 2 (Pensiones y Otras Prestaciones Sociales) | 0.012 |
| tejgfun\_f5ct05vivpc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados), en la Categoría 5 (Gastos Corrientes) para la Función 19 (Vivienda y Desarrollo Urbano) | 0.012 |
| tejgrb\_impmpc | Total Ejecutado per cápita en el Rubro de Impuestos Municipales | 0.011 |
| \_tejgge\_r09ct05biser | Total Ejecutado rezagado en el Rubro 09 (Recursos Directamente Recaudados), Categoría 5 (Gastos Corrientes), para la genérica de Bienes y Servicios | 0.011 |
| tejgfun\_f2ct05ambpc | Total Ejecutado de la Fuente 2 (Recursos Directamente Recaudados), en la Categoría 5 (Gastos Corrientes) para la Función 17 ( Ambiente) | 0.01 |
| tejgfun\_f5ct06viv | Total Ejecutado de la Fuente 5 (Recursos Determinados), en la Categoría 6 (Gastos de Capital) para la Función 19 (Vivienda y Desarrollo Urbano) | 0.009 |
| tejgtotfun\_f5pgrco | Total Ejecutado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 03 (Planeamiento, Gestión y Reserva de Contingencia) | 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 2.** Correlación entre las 20 variables más importantes según el criterio de impureza de Gini y las variables de Canon.

Calendario

Descripción generada automáticamente

**Variables más importantes según el criterio de SHAP Values**

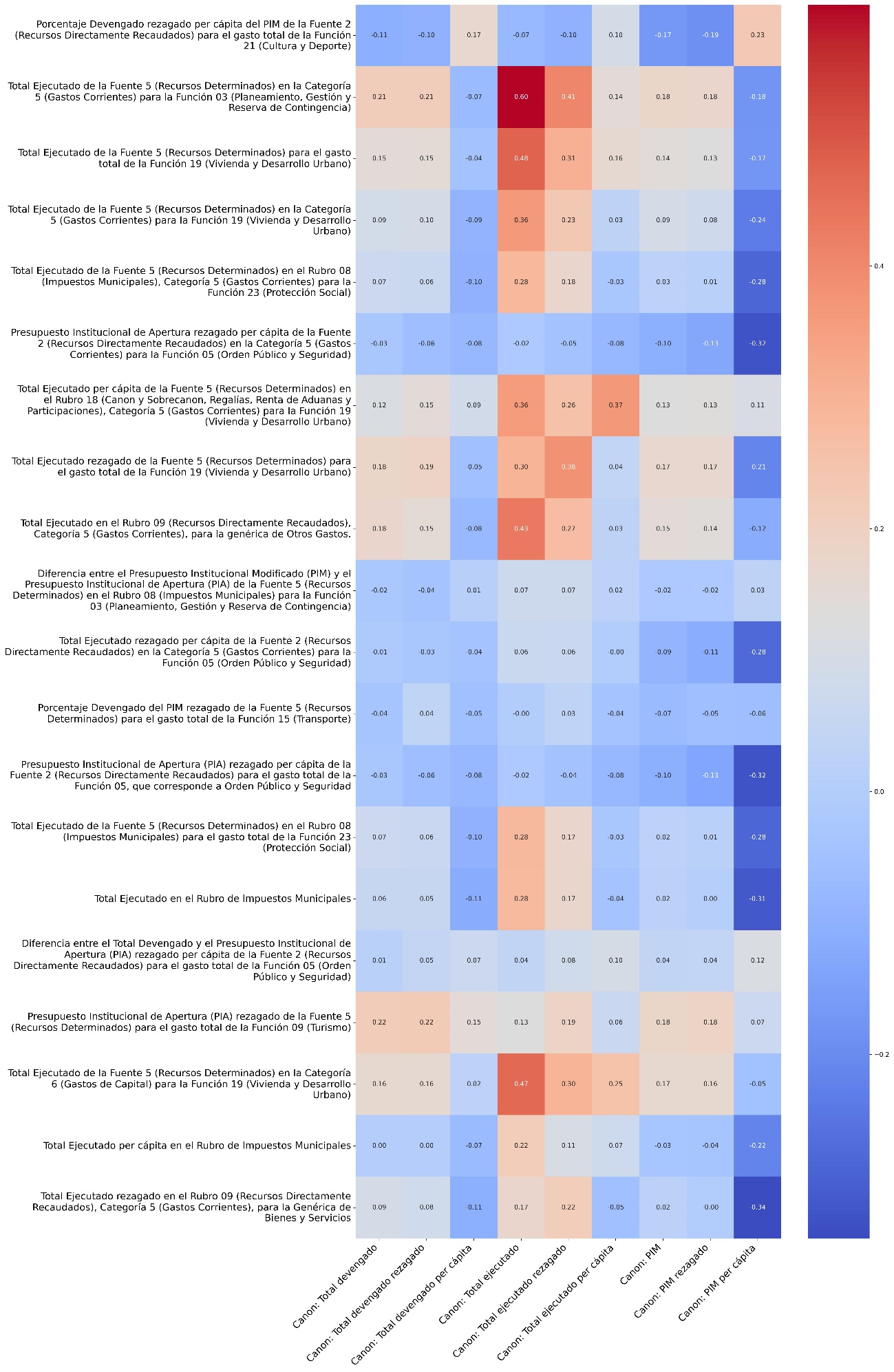
En esta sección se presentan las 20 variables más importantes según el criterio de SHAP Values para el modelo óptimo.

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

|  |  |  |
| --- | --- | --- |
| **Variable** | **Etiqueta** | **SHAP Values** |
| \_devppimtotfun\_f2cydeppc | Porcentaje Devengado rezagado per cápita del PIM de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función 21 (Cultura y Deporte) | 0.055 |
| 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) | 0.016 |
| tejgtotfun\_f5viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 19 (Vivienda y Desarrollo Urbano) | 0.016 |
| tejgfun\_f5ct05viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función 19 (Vivienda y Desarrollo Urbano) | 0.011 |
| tejgfun\_f5r08ct05prots | Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales), Categoría 5 (Gastos Corrientes) para la Función 23 (Protección Social) | 0.01 |
| \_piagfun\_f2ct05opsegpc | Presupuesto Institucional de Apertura rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad) | 0.01 |
| 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) | 0.008 |
| \_tejgtotfun\_f5viv | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 19 (Vivienda y Desarrollo Urbano) | 0.008 |
| tejgge\_r09ct05otgst | Total Ejecutado en el Rubro 09 (Recursos Directamente Recaudados), Categoría 5 (Gastos Corrientes), para la genérica de Otros Gastos. | 0.007 |
| dfgpimpiatotfun\_f5r08pgrco | Diferencia entre el Presupuesto Institucional Modificado (PIM) y el Presupuesto Institucional de Apertura (PIA) de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para la Función 03 (Planeamiento, Gestión y Reserva de Contingencia) | 0.006 |
| \_tejgfun\_f2ct05opsegpc | Total Ejecutado rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función 05 (Orden Público y Seguridad) | 0.006 |
| \_devppimtotfun\_f5trans | Porcentaje Devengado del PIM rezagado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 15 (Transporte) | 0.005 |
| \_piagtotfun\_f2opsegpc | Presupuesto Institucional de Apertura (PIA) rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función 05, que corresponde a Orden Público y Seguridad | 0.004 |
| tejgtotfun\_f5r08prots | Total Ejecutado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para el gasto total de la Función 23 (Protección Social) | 0.004 |
| tejgrb\_impm | Total Ejecutado en el Rubro de Impuestos Municipales | 0.004 |
| \_dfgdevpiagtotfun\_f2opsegpc | Diferencia entre el Total Devengado y el Presupuesto Institucional de Apertura (PIA) rezagado per cápita de la Fuente 2 (Recursos Directamente Recaudados) para el gasto total de la Función 05 (Orden Público y Seguridad) | 0.003 |
| \_piagtotfun\_f5turi | Presupuesto Institucional de Apertura (PIA) rezagado de la Fuente 5 (Recursos Determinados) para el gasto total de la Función 09 (Turismo) | 0.003 |
| tejgfun\_f5ct06viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función 19 (Vivienda y Desarrollo Urbano) | 0.003 |
| tejgrb\_impmpc | Total Ejecutado per cápita en el Rubro de Impuestos Municipales | 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 | 0.003 |

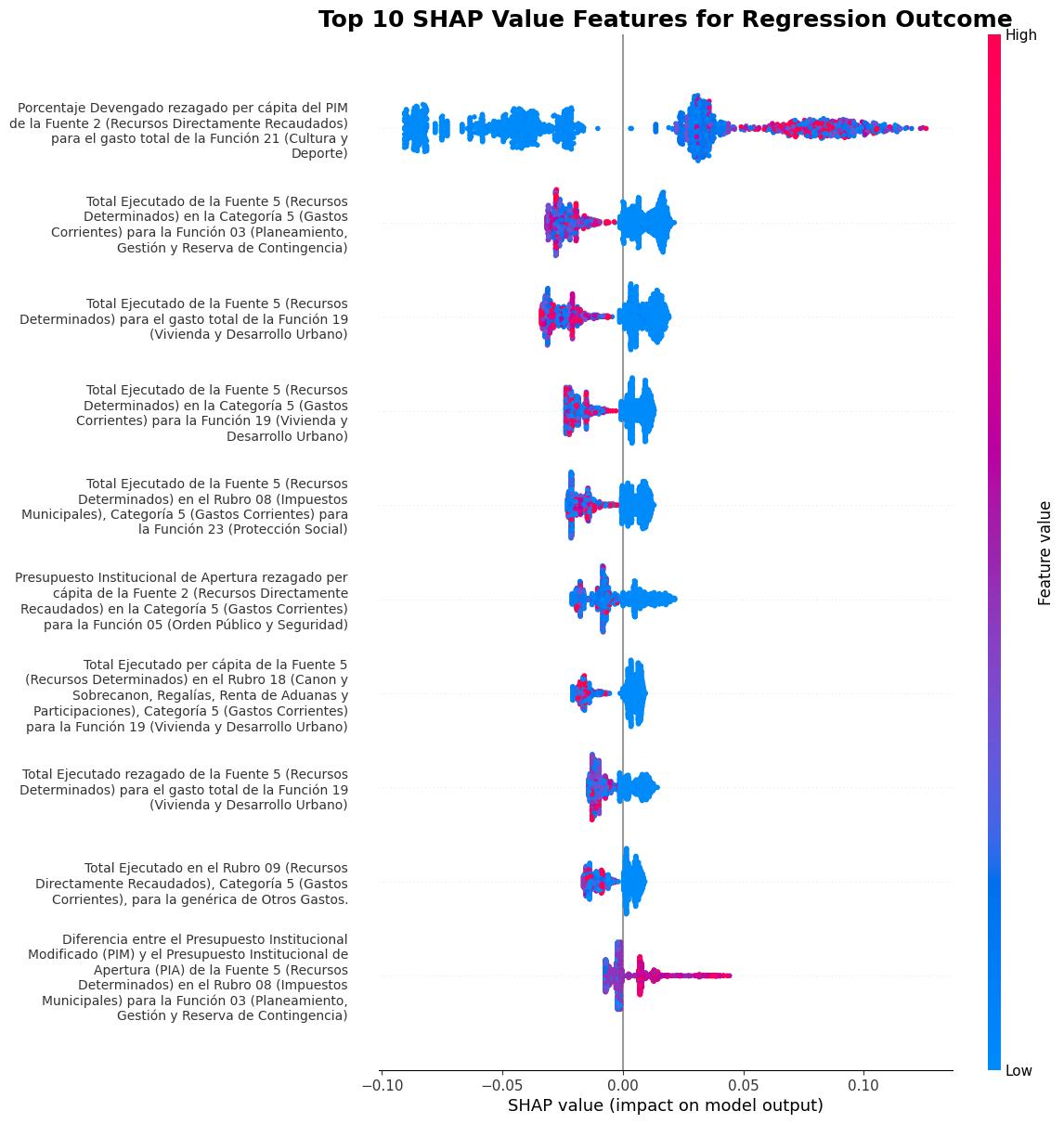
Asimismo, se visualiza la correlación entre las 20 variables más importantes según el criterio SHAP Values y las variables de Canon.

**Gráfico 3.** 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 4.** 10 variables más importantes según criterio SHAP para casos positivos de Corrupción Amplia



**ANEXOS**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | **n\_estimators** | **max\_depth** | **max\_features** |
| Regression Forest | 1000 | 20 | - |
| LGBM Classifier | 500 | 2 | - |
| Gradient Boosting Trees | 500 | 2 | 20% |
| Random Forest | 250 | 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 | 100 |
| Ridge | 100 |
| Elastic Net | 100 |
| Fuente: elaboración propia | |

**RESULTADOS CON UN STEPS DE 0.01 PARA EL MODELO REGRESSION FOREST**

La siguiente tabla presenta los resultados óptimos cuando se consideran steps de 0.01 unidades para establecer el threshold en el modelo Regression. Se muestra que el modelo óptimo, según la métrica F1, es el modelo Regression Forest entrenado con el conjunto de entrenamiento SMOTE con un threshold de 0.45.

**Tabla 95**. 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 = 0.45)** |
| F1 | 0.506 | 0.541 | 0.541 | 0.541 | 0.551 | 0.522 | 0.523 | 0.662 |
| Accuracy | 0.833 | 0.780 | 0.78 | 0.78 | 0.896 | 0.906 | 0.908 | 0.879 |
| AUC ROC | 0.556 | 0.636 | 0.636 | 0.636 | 0.808 | 0.778 | 0.741 | 0.814 |
| F1 (Sí) | 0.908 | 0.872 | 0.872 | 0.872 | 0.945 | 0.950 | 0.952 | 0.933 |
| F1 (No) | 0.104 | 0.209 | 0.209 | 0.209 | 0.157 | 0.093 | 0.095 | 0.390 |
| Fuente: elaboración propia | | | | | | | | |