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

**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 intensa
* 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 intensa** | **Total de observaciones en las que no ocurrió corrupción intensa** |
| Original | 964 | 692 | 272 |
| SMOTE | 1 384 | 692 | 692 |
| SMOTE Tomek-Links | 1 368 | 684 | 684 |
| Naive Random Oversampling | 1 384 | 692 | 692 |
| 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 | 250, 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 **Random Forest** entrenado con el conjunto de entrenamiento SMOTE-Tomek Links. 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-Tomek Links

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Métrica** | **Regresión Logística** | **Lasso** | **Ridge** | **Elastic Net** | **Random Forest** | **Gradient Boosting Trees** | **LGBM Classifier** | **Regression Forest** |
| F1 | 0.563 | 0.568 | 0.568 | 0.568 | 0.624 | 0.589 | 0.605 | 0.594 |
| Accuracy | 0.626 | 0.626 | 0.626 | 0.626 | 0.725 | 0.705 | 0.717 | 0.638 |
| AUC ROC | 0.562 | 0.604 | 0.604 | 0.604 | 0.639 | 0.619 | 0.629 | 0.644 |
| F1 (Sí) | 0.729 | 0.726 | 0.726 | 0.726 | 0.818 | 0.808 | 0.816 | 0.727 |
| F1 (No) | 0.397 | 0.411 | 0.411 | 0.411 | 0.430 | 0.371 | 0.394 | 0.460 |
| Fuente: elaboración propia | | | | | | | | |

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

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**Grafico 1.** Grid Search Accuracy vs. Grid Search Parameters

**Variables más importantes según el criterio de impureza de Gini**

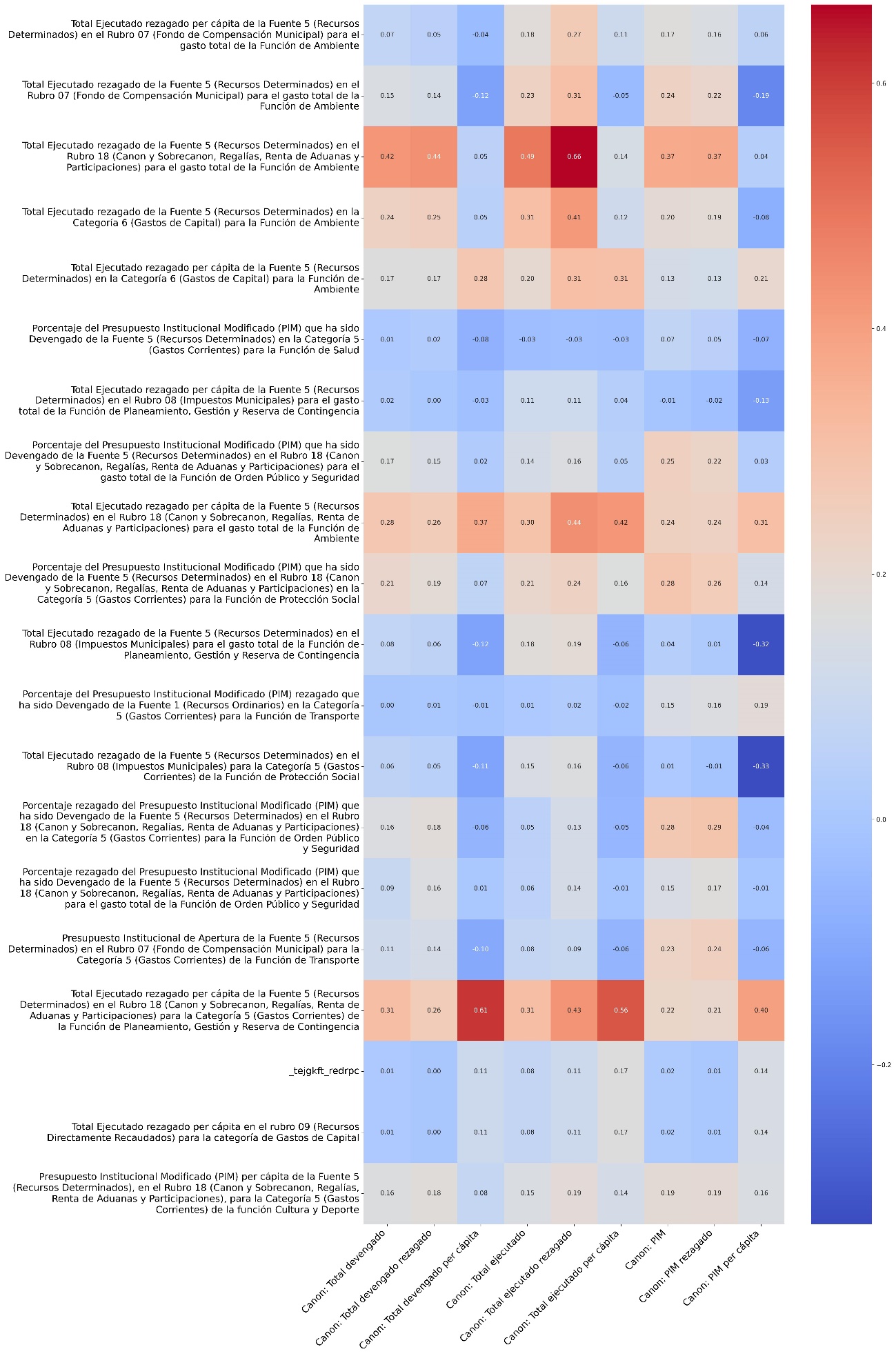
En esta sección se presentan las 20 variables más importantes según el criterio de impureza de Gini (estimado mediante el comando feature importance) para el modelo óptimo.

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

|  |  |  |
| --- | --- | --- |
| **Variable** | **Eiqueta** | **Importance Score** |
| \_tejgtotfun\_f5r07ambpc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para el gasto total de la Función de Ambiente. | 0.045 |
| \_tejgtotfun\_f5r07amb | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para el gasto total de la Función de Ambiente. | 0.015 |
| \_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 el gasto total de la Función de Ambiente | 0.013 |
| \_tejgfun\_f5ct06amb | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función de Ambiente. | 0.010 |
| \_tejgfun\_f5ct06ambpc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función de Ambiente. | 0.009 |
| devppimfun\_f5ct05salud | Porcentaje del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función de Salud. | 0.008 |
| \_tejgtotfun\_f5r08pgrcopc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para el gasto total de la Función de Planeamiento, Gestión y Reserva de Contingencia. | 0.007 |
| devppimtotfun\_f5r18opseg | Porcentaje del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para el gasto total de la Función de Orden Público y Seguridad. | 0.007 |
| \_tejgtotfun\_f5r18ambpc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para el gasto total de la Función de Ambiente. | 0.006 |
| devppimfun\_f5r18ct05prots | Porcentaje del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) en la Categoría 5 (Gastos Corrientes) para la Función de Protección Social. | 0.006 |
| \_tejgtotfun\_f5r08pgrco | Total Ejecutado rezagadode la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para el gasto total de la Función de Planeamiento, Gestión y Reserva de Contingencia. | 0.005 |
| \_devppimfun\_f1ct05trans | Porcentaje del Presupuesto Institucional Modificado (PIM) rezagado que ha sido Devengado de la Fuente 1 (Recursos Ordinarios) en la Categoría 5 (Gastos Corrientes) para la Función de Transporte. | 0.005 |
| \_tejgfun\_f5r08ct05prots | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para la Categoría 5 (Gastos Corrientes) de la Función de Protección Social | 0.005 |
| \_devppimfun\_f5r18ct05opseg | Porcentaje rezagado del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) en la Categoría 5 (Gastos Corrientes) para la Función de Orden Público y Seguridad. | 0.004 |
| \_devppimtotfun\_f5r18opseg | Porcentaje rezagado del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para el gasto total de la Función de Orden Público y Seguridad. | 0.004 |
| piagfun\_f5r07ct05trans | Presupuesto Institucional de Apertura de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para la Categoría 5 (Gastos Corrientes) de la Función de Transporte. | 0.004 |
| \_tejgfun\_f5r18ct05pgrcopc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Categoría 5 (Gastos Corrientes) de la Función de Planeamiento, Gestión y Reserva de Contingencia. | 0.004 |
| \_tejgkft\_redrpc | tejgkft\_redrpc | 0.004 |
| \_tejgct\_r09gstcppc | Total Ejecutado rezagado per cápita en el rubro 09 (Recursos Directamente Recaudados) para la categoría de Gastos de Capital | 0.003 |
| pimgfun\_f5r18ct05cydeppc | Presupuesto Institucional Modificado (PIM) per cápita de la Fuente 5 (Recursos Determinados), en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones), para la Categoría 5 (Gastos Corrientes) de la función Cultura y Deporte | 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 2.** 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**

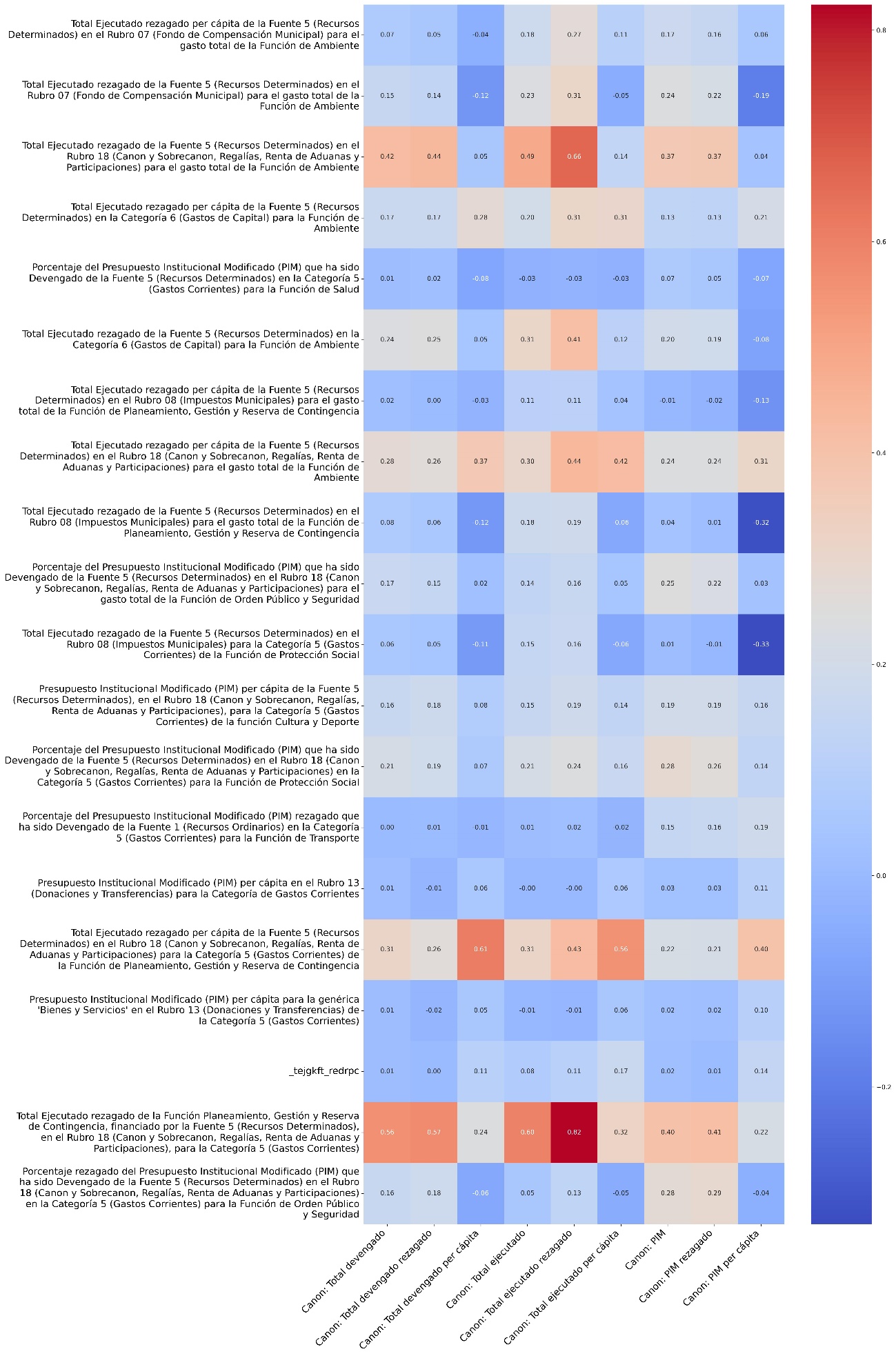
En esta sección se presentan las 20 variables más importantes según el criterio de SHAP Values para el modelo óptimo.

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

|  |  |  |
| --- | --- | --- |
| **Variable** | **Etiqueta** | **SHAP Values** |
| \_tejgtotfun\_f5r07ambpc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para el gasto total de la Función de Ambiente. | 0.045 |
| \_tejgtotfun\_f5r07amb | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) para el gasto total de la Función de Ambiente. | 0.015 |
| \_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 el gasto total de la Función de Ambiente | 0.014 |
| \_tejgfun\_f5ct06ambpc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función de Ambiente. | 0.009 |
| devppimfun\_f5ct05salud | Porcentaje del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función de Salud. | 0.009 |
| \_tejgfun\_f5ct06amb | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función de Ambiente. | 0.009 |
| \_tejgtotfun\_f5r08pgrcopc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para el gasto total de la Función de Planeamiento, Gestión y Reserva de Contingencia. | 0.007 |
| \_tejgtotfun\_f5r18ambpc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para el gasto total de la Función de Ambiente. | 0.006 |
| \_tejgtotfun\_f5r08pgrco | Total Ejecutado rezagadode la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para el gasto total de la Función de Planeamiento, Gestión y Reserva de Contingencia. | 0.006 |
| devppimtotfun\_f5r18opseg | Porcentaje del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para el gasto total de la Función de Orden Público y Seguridad. | 0.005 |
| \_tejgfun\_f5r08ct05prots | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para la Categoría 5 (Gastos Corrientes) de la Función de Protección Social | 0.005 |
| pimgfun\_f5r18ct05cydeppc | Presupuesto Institucional Modificado (PIM) per cápita de la Fuente 5 (Recursos Determinados), en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones), para la Categoría 5 (Gastos Corrientes) de la función Cultura y Deporte | 0.005 |
| devppimfun\_f5r18ct05prots | Porcentaje del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) en la Categoría 5 (Gastos Corrientes) para la Función de Protección Social. | 0.005 |
| \_devppimfun\_f1ct05trans | Porcentaje del Presupuesto Institucional Modificado (PIM) rezagado que ha sido Devengado de la Fuente 1 (Recursos Ordinarios) en la Categoría 5 (Gastos Corrientes) para la Función de Transporte. | 0.004 |
| pimgct\_r13gstcrpc | Presupuesto Institucional Modificado (PIM) per cápita en el Rubro 13 (Donaciones y Transferencias) para la Categoría de Gastos Corrientes. | 0.004 |
| \_tejgfun\_f5r18ct05pgrcopc | Total Ejecutado rezagado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Categoría 5 (Gastos Corrientes) de la Función de Planeamiento, Gestión y Reserva de Contingencia. | 0.004 |
| pimgge\_r13ct05biserpc | Presupuesto Institucional Modificado (PIM) per cápita para la genérica 'Bienes y Servicios' en el Rubro 13 (Donaciones y Transferencias) de la Categoría 5 (Gastos Corrientes). | 0.004 |
| \_tejgkft\_redrpc | \_tejgkft\_redrpc | 0.004 |
| \_tejgfun\_f5r18ct05pgrco | Total Ejecutado rezagado de la Función Planeamiento, Gestión y Reserva de Contingencia, financiado por la Fuente 5 (Recursos Determinados), en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones), para la Categoría 5 (Gastos Corrientes). | 0.004 |
| \_devppimfun\_f5r18ct05opseg | Porcentaje rezagado del Presupuesto Institucional Modificado (PIM) que ha sido Devengado de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) en la Categoría 5 (Gastos Corrientes) para la Función de Orden Público y Seguridad. | 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 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 4.** 10 variables más importantes según criterio SHAP para casos positivos de Corrupción intensa

Imagen que contiene Gráfico

Descripción generada automáticamente

**ANEXOS**

**Anexo 1. Combinación óptima de hiperparámetros para modelos SMOTE Tomek-Links basados 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 | 250 | 30 | 40% |
| Fuente: elaboración propia | | | |

**Anexo 2. Combinación óptima de hiperparámetros para modelos SMOTE Tomek-Links 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 **original** con un threshold de 0.65.

**Tabla 95**. Métricas de desempeño de los modelos entrenados con el conjunto de entrenamiento Original

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Métrica** | **Regresión Logística** | **Lasso** | **Ridge** | **Elastic Net** | **Random Forest** | **Gradient Boosting Trees** | **LGBM Classifier** | **Regression Forest (threshold = 0.65)** |
| F1 | 0.550 | 0.600 | 0.600 | 0.600 | 0.566 | 0.554 | 0.577 | 0.633 |
| Accuracy | 0.643 | 0.722 | 0.722 | 0.722 | 0.729 | 0.703 | 0.744 | 0.708 |
| AUC ROC | 0.561 | 0.625 | 0.625 | 0.625 | 0.656 | 0.621 | 0.644 | 0.664 |
| F1 (Sí) | 0.754 | 0.821 | 0.821 | 0.821 | 0.832 | 0.812 | 0.843 | 0.799 |
| F1 (No) | 0.345 | 0.378 | 0.378 | 0.378 | 0.300 | 0.297 | 0.312 | 0.467 |
| Fuente: elaboración propia | | | | | | | | |