**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
* Regression Forest Threshold: 0.05

**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 108 a 10-6, 100 valores en escala logarítmica |
| Ridge | De 108 a 10-6, 100 valores en escala logarítmica |
| Elastic Net | De 108 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 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 cuando estos fueron entrenados con el conjunto de entrenamiento SMOTE Tomek-Links. Tomando en cuenta la métrica F1, el modelo con el mejor desempeño es el modelo **Regression Forest** entrenado con el conjunto de entrenamiento SMOTE Tomek-Links a un threshold de 0.55.

**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 (threshold = 0.55)** |
| F1 | 0.486 | 0.407 | 0.407 | 0.407 | 0.565 | 0.563 | 0.525 | 0.666 |
| Accuracy | 0.677 | 0.494 | 0.494 | 0.494 | 0.911 | 0.908 | 0.911 | 0.872 |
| AUC ROC | 0.554 | 0.630 | 0.630 | 0.630 | 0.757 | 0.727 | 0.699 | 0.770 |
| F1 (Sí) | 0.799 | 0.634 | 0.634 | 0.634 | 0.953 | 0.952 | 0.953 | 0.928 |
| F1 (No) | 0.173 | 0.180 | 0.180 | 0.180 | 0.178 | 0.174 | 0.098 | 0.404 |
| Fuente: elaboración propia | | | | | | | | |

Asimismo, se presentan los 10 modelos con mejor desempeño 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**. 10 modelos con mayor poder predictivo considerando los distintos métodos de remuestreo

|  |  |
| --- | --- |
| **Modelo** | **F1** |
| Regression Forest ST (threshold = 0.55) | 0.666 |
| Regression Forest NRO (threshold = 0.55) | 0.650 |
| Regression Forest NRO (threshold = 0.50 | 0.648 |
| Regression Forest ST (threshold = 0.50) | 0.647 |
| Regression Forest S (threshold = 0.55) | 0.645 |
| Regression Forest S (threshold = 0.60) | 0.635 |
| Regression Forest NRO (threshold = 0.60) | 0.634 |
| Regression Forest S (threshold = 0.65) | 0.625 |
| Regression Forest O (threshold = 0.90) | 0.621 |
| Regression Forest ST (threshold = 0.60) | 0.621 |
| Fuente: elaboración propia | |

El gráfico 1. Muestra el ajuste del modelo óptimo (**Regression Forest** entrenado con el conjunto de SMOTE Tomek-Links) 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 SMOTE Tomek-Links 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] con pasos de 0.05.

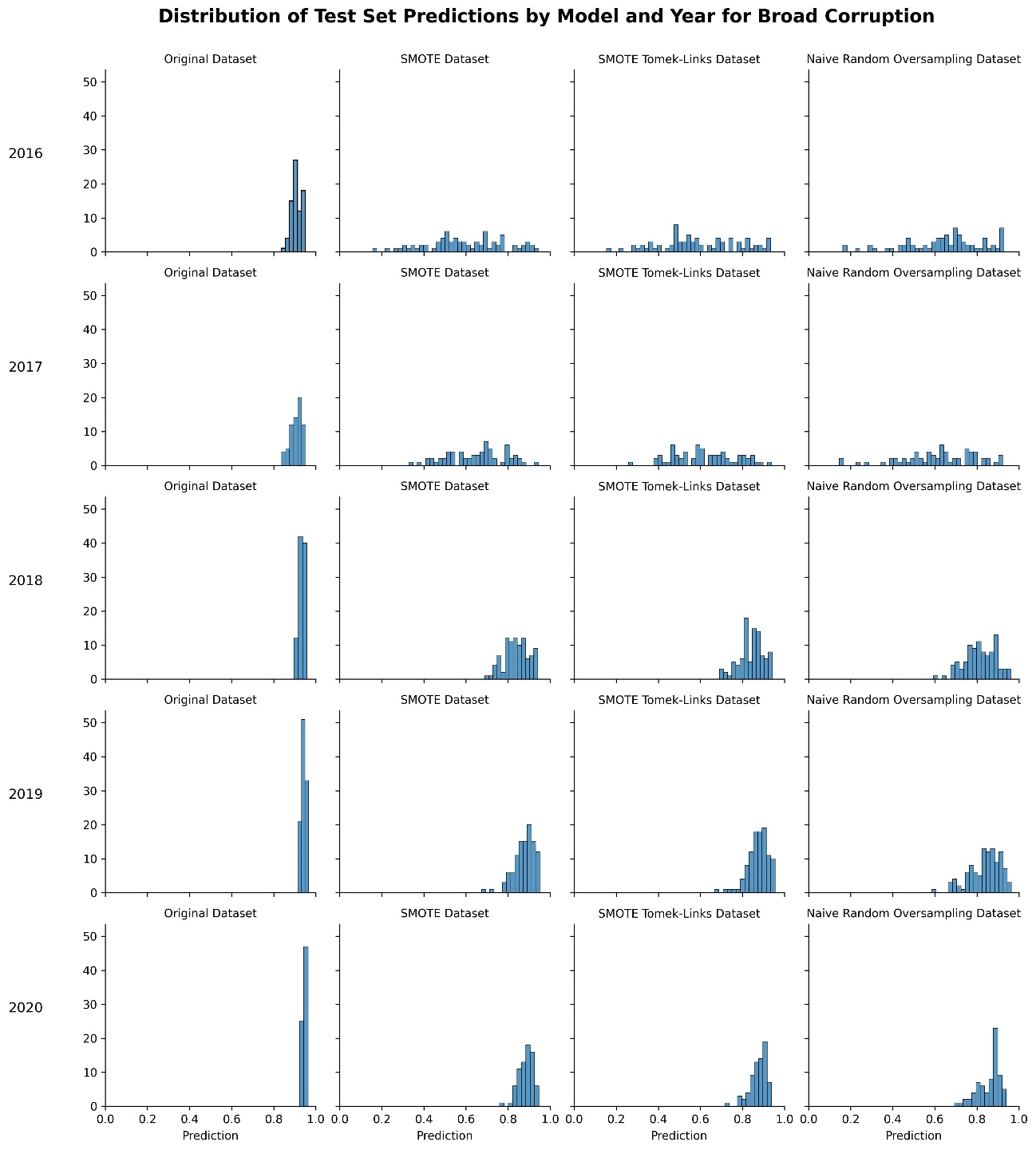
**Gráfico 2.** F1 en función de distintos thesholds para distintos conjuntos de entrenamiento.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**Distribución de predicciones**

El gráfico 3 presenta la distribución de predicciones para distintos años en función de los distintos métodos de remuestreo empleados. Las predicciones se realizan sobre la prueba set. El modelo óptimo es aquel que fue entrenado con el conjunto de entrenamiento SMOTE Tomek-Links, a un threshold de 0.55.



**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\_f5ct05vivpc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.159 |
| tejgfun\_f5ct05viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.134 |
| tejgfun\_f5r18ct05viv | Total Ejecutado 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 Vivienda y Desarrollo Urbano | SIAF | 0.103 |
| 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) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.084 |
| tejgtotfun\_f5r18cydeppc | Total Ejecutado de gasto total per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Cultura y Deporte | SIAF | 0.076 |
| tejgtotfun\_f5r18cydep | Total Ejecutado de gasto total de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Cultura y Deporte | SIAF | 0.048 |
| tejgtotfun\_f5r18viv | Total Ejecutado de gasto total de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.046 |
| tejgtotfun\_f5r18vivpc | Total Ejecutado de gasto total per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.042 |
| tejgfun\_f5r08ct05ambpc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) en la Categoría 5 (Gastos Corrientes) para la Función Ambiente | SIAF | 0.022 |
| \_tejgfun\_f5r07ct06viv | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) en la Categoría 6 (Gastos de Capital) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.021 |
| tejgfun\_f5ct06cydeppc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Cultura y Deporte | SIAF | 0.018 |
| tejgfun\_f2ct05vivpc | Total Ejecutado per cápita de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.016 |
| tejgge\_r08ct05pobsopc | Total Ejecutado per cápita del Rubro 08 (Impuestos Municipales) en la Categoría 5 (Gastos Corrientes) para la Genérica Personal y Obligaciones Sociales | SIAF | 0.011 |
| tejgtotfun\_f5r08ambpc | Total Ejecutado de gasto total per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) para la Función Ambiente | SIAF | 0.010 |
| tejgge\_r08ct05pobso | Total Ejecutado del Rubro 08 (Impuestos Municipales) en la Categoría 5 (Gastos Corrientes) para la Genérica Personal y Obligaciones Sociales | SIAF | 0.009 |
| tejgfun\_f2ct05viv | Total Ejecutado de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.007 |
| tejgfun\_f5ct06cydep | Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Cultura y Deporte | SIAF | 0.006 |
| mod\_28 | La municipalidad dispone de otros instrumentos de gestión (dicotómica) | Renamu | 0.006 |
| tejgtotfun\_f5viv | Total Ejecutado de gasto total de la Fuente 5 (Recursos Determinados) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.005 |
| tejgge\_r08ct05biser | Total Ejecutado del Rubro 08 (Impuestos Municipales) en la Categoría 5 (Gastos Corrientes) para la Genérica Bienes y Servicios | SIAF | 0.004 |

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.

Gráfico, Calendario

Descripción generada automáticamente con confianza media

**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\_f5ct05vivpc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.049 |
| tejgfun\_f5ct05viv | Total Ejecutado de la Fuente 5 (Recursos Determinados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.041 |
| tejgfun\_f5r18ct05viv | Total Ejecutado 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 Vivienda y Desarrollo Urbano | SIAF | 0.030 |
| 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) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.025 |
| tejgtotfun\_f5r18cydeppc | Total Ejecutado de gasto total per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Cultura y Deporte | SIAF | 0.022 |
| tejgtotfun\_f5r18cydep | Total Ejecutado de gasto total de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Cultura y Deporte | SIAF | 0.014 |
| tejgtotfun\_f5r18viv | Total Ejecutado de gasto total de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.014 |
| tejgtotfun\_f5r18vivpc | Total Ejecutado de gasto total per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 18 (Canon y Sobrecanon, Regalías, Renta de Aduanas y Participaciones) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.012 |
| \_tejgfun\_f5r07ct06viv | Total Ejecutado rezagado de la Fuente 5 (Recursos Determinados) en el Rubro 07 (Fondo de Compensación Municipal) en la Categoría 6 (Gastos de Capital) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.011 |
| mod\_28 | La municipalidad dispone de otros instrumentos de gestión (dicotómica) | Renamu | 0.008 |
| tejgfun\_f5r08ct05ambpc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en el Rubro 08 (Impuestos Municipales) en la Categoría 5 (Gastos Corrientes) para la Función Ambiente | SIAF | 0.006 |
| \_piagfun\_f5ct06turipc | Presupuesto Institucional de Apertura rezagado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Turismo | SIAF | 0.006 |
| tejgtotfun\_f5turi | Total Ejecutado de gasto total de la Fuente 5 (Recursos Determinados) para la Función Turismo | SIAF | 0.005 |
| mod\_18 | La municipalidad dispone de un Plan Local de Igualdad de Género (dicotómica) | Renamu | 0.005 |
| tejgfun\_f5ct06cydeppc | Total Ejecutado per cápita de la Fuente 5 (Recursos Determinados) en la Categoría 6 (Gastos de Capital) para la Función Cultura y Deporte | SIAF | 0.005 |
| tejgfun\_f2ct05vivpc | Total Ejecutado per cápita de la Fuente 2 (Recursos Directamente Recaudados) en la Categoría 5 (Gastos Corrientes) para la Función Vivienda y Desarrollo Urbano | SIAF | 0.005 |
| \_dfgpimpiafun\_f5r07ct06turi | Diferencia rezagada entre el PIM y el PIA de la Fuente 5 (Recursos Determinados), Rubro 07 (Fondo de Compensación Municipal) en la Categoría 6 (Gastos de Capital) para la Función Turismo | SIAF | 0.004 |
| devppimtotfun\_f5turi | Porcentaje Devengado del PIM de gasto total de la Fuente 5 (Recursos Determinados) para la Función Turismo | SIAF | 0.004 |
| tejgfun\_f5r07ct05trans | Total Ejecutado de la Fuente 5 (Recursos Determinados), Rubro 07 (Fondo de Compensación Municipal) en la Categoría 5 (Gastos Corrientes) para la Función Transporte | SIAF | 0.003 |
| \_devppimgge\_r07ct05dotra | Porcentaje Devengado del PIM rezagado del Rubro 07 (Fondo de Compensación Municipal) en la Categoría 5 (Gastos Corrientes) para la Genérica Donaciones y Transferencias | 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.Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

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

Gráfico

Descripción generada automáticamente

**ANEXOS**

**Anexo 1. Combinación óptima de hiperparámetros para modelos SMOTE Tomek-Links en árboles**

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

**Anexo 2. Combinación óptima de hiperparámetros para modelos SMOTE Tomek-Links de regularización**

|  |  |
| --- | --- |
| **Modelo** | **Cs (Fuerza de la regularización)** |
| Lasso | 1450828.778 |
| Ridge | 1450828.778 |
| Elastic Net | 1450828.778 |
| Fuente: elaboración propia | |