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

**Detalles sobre el presente reporte**

* Fecha: 03 de enero 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: 2007-2020
* Tipo de predicción: clasificación

**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 | 687 | 277 |
| SMOTE | 1 374 | 687 | 687 |
| SMOTE Tomek-Links | 1 354 | 677 | 677 |
| Naive Random Oversampling | 1 474 | 687 | 687 |
| 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% |
| 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** |
| F1 | 0.546 | 0.596 | 0.596 | 0.596 | 0.650 | 0.614 |
| Accuracy | 0.604 | 0.650 | 0.650 | 0.650 | 0.749 | 0.725 |
| AUC ROC | 0.582 | 0.635 | 0.635 | 0.635 | 0.712 | 0.672 |
| F1 (Sí) | 0.708 | 0.743 | 0.743 | 0.743 | 0.836 | 0.821 |
| F1 (No) | 0.383 | 0.449 | 0.449 | 0.449 | 0.464 | 0.406 |
| Fuente: elaboración propia | | | | | | |

**Variables más importantes**

En esta sección se presentan las 20 variables más importantes según el criterio de feature importance para el modelo óptimo (Random Forest entrenado con el conjunto de entrenamiento SMOTE-Tomek Links).

Tabla 7. Veinte variables más importantes de acuerdo con el criterio de features importance

|  |  |
| --- | --- |
| **Variable** | **Importance Score** |
| \_tejgfun\_f5ct06ambpc | 0.023 |
| \_tejgfun\_f5ct06amb | 0.015 |
| \_tejgfun\_f5ct06vivpc | 0.007 |
| \_tejgfun\_f5r18ct05pgrco | 0.006 |
| devppimfun\_f5r18ct05agro | 0.006 |
| \_tejgfun\_f5ct06viv | 0.005 |
| \_piagtotfun\_f5r07energ | 0.005 |
| \_tejgtotfun\_f2opsegpc | 0.005 |
| piagtotfun\_f5r18cydeppc | 0.004 |
| tejgtotfun\_f5r18opseg | 0.004 |
| \_devppimfun\_f5r18ct05salud | 0.004 |
| devppimtotfun\_f5r18opseg | 0.004 |
| \_tejgtotfun\_f2opseg | 0.003 |
| tejgfun\_f5ct06vivpc | 0.003 |
| tejgfun\_f5r18ct05cydeppc | 0.003 |
| \_piagtotfun\_f5r07energpc | 0.003 |
| piagfun\_f5r18ct06cydeppc | 0.003 |
| dfgdevpiagfun\_f5r18ct06cydeppc | 0.003 |
| \_tejgfun\_f5ct06opsegpc | 0.003 |