**REPORTE DE RESULTADOS**

**Detalles sobre el presente reporte**

* 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 de análisis: 2007-2020
* 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 variabilidad**. Se descartaron todas aquellas variables constantes, es decir, con una variabilidad de 0.
3. **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%.
4. **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 | 17 539 | 14740 |
| 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 | 1 443 | 514 | 929 |
| SMOTE | 1 853 | 929 | 929 |
| SMOTE Tomek-Links | 1 808 | 904 | 904 |
| Naive Random Oversampling | 1 858 | 929 | 929 |
| 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 Extreme Gradient Boosting (XGB) 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** |
| F1 | 0.568 | 0.588 | 0.588 | 0.588 | 0.615 | 0.625 |
| Accuracy | 0.586 | 0.599 | 0.599 | 0.599 | 0.648 | 0.662 |
| AUC ROC | 0.599 | 0.646 | 0.646 | 0.646 | 0.654 | 0.642 |
| F1 (Sí) | 0.657 | 0.656 | 0.656 | 0.656 | 0.728 | 0.744 |
| F1 (No) | 0.480 | 0.521 | 0.521 | 0.521 | 0.502 | 0.506 |
| 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 (Extreme Gradient Boosting entrenado con el conjunto de entrenamiento SMOTE).

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

|  |  |
| --- | --- |
| **Variable** | **Importance Score** |
| \_tejgfun\_f5ct06amb | 0.031 |
| tejgtotfun\_f5opsegpc | 0.031 |
| \_tejgtotfun\_f5r08pgrco | 0.020 |
| tejgfun\_f5ct06cydep | 0.017 |
| tejgtotfun\_f2opsegpc | 0.012 |
| piagfun\_f1ct05pgrco | 0.011 |
| \_tejgfun\_f5r18ct06cydeppc | 0.009 |
| tdvgtotfun\_f4cydeppc | 0.008 |
| piagfun\_f5ct06ambpc | 0.008 |
| tejgge\_r08ct05biserpc | 0.008 |
| \_piagtotfun\_f5r18ambpc | 0.008 |
| devppimtotfun\_f1pgrco | 0.007 |
| tejgft\_redr | 0.007 |
| \_tejgfun\_f5r07ct05trans | 0.006 |
| pimgfun\_f5r07ct05saludpc | 0.006 |
| \_devppimfun\_f5ct05amb | 0.006 |
| \_dfgpimpiatotfun\_f5r18opseg | 0.005 |
| \_tdvgfun\_f5r07ct05cydeppc | 0.005 |
| \_tdvgct\_r13gstcr | 0.005 |

**Mapas de predicción de dos categorías**

Los siguientes gráficos presentan los mapas de predicción según el modelo óptimo (Extreme Gradient Boosting entrenado con el conjunto de entrenamiento SMOTE) comparando tres categorías: la corrupción real, la corrupción predicha, y los 100 distritos con mayor probabilidad de que ocurra corrupción por año.

Gráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteMapa, Gráfico de dispersión

Descripción generada automáticamente con confianza mediaGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico

Descripción generada automáticamente con confianza bajaGráfico, Mapa, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Mapa

Descripción generada automáticamente con confianza mediaMapa

Descripción generada automáticamenteGráfico, Mapa

Descripción generada automáticamente con confianza media

**ANEXOS**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | **n\_estimators** | **max\_depth** | **max\_features** |
| Gradient Boosting Trees | 1000 | 2 | 20% |
| Random Forest | 250 | 30 | 20% |
| Fuente: elaboración propia | | | |

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

|  |  |
| --- | --- |
| **Modelo** | **Cs (Fuerza de la regularización)** |
| Lasso | 10 |
| Ridge | 10 |
| Elastic Net | 10 |
| Fuente: elaboración propia | |