**REPORTE DE RESULTADOS**

**Detalles sobre el presente reporte**

* Nivel de observación: año inicial del reporte por municipalidad y año.
* Variable predicha: corrupción intensa
* Periodo de análisis: 2016-2020
* 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 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 intensa** | **Total de observaciones en las que no ocurrió corrupción intensa** |
| Original | 967 | 284 | 683 |
| SMOTE | 1 366 | 683 | 683 |
| SMOTE Tomek-Links | 1 212 | 606 | 606 |
| Naive Random Oversampling | 1 366 | 683 | 683 |
| 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 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 NRO

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Métrica** | **Regresión Logística** | **Lasso** | **Ridge** | **Elastic Net** | **Random Forest** | **Gradient Boosting Trees** |
| F1 | 0.285 | 0.348 | 0.348 | 0.348 | 0.647 | 0.595 |
| Accuracy | 0.299 | 0.354 | 0.354 | 0.354 | 0.754 | 0.733 |
| AUC ROC | 0.546 | 0.554 | 0.555 | 0.554 | 0.684 | 0.670 |
| F1 (Sí) | 0.185 | 0.287 | 0.287 | 0.287 | 0.842 | 0.831 |
| F1 (No) | 0.385 | 0.410 | 0.410 | 0.410 | 0.452 | 0.358 |
| Fuente: elaboración propia | | | | | | |

Debe tenerse en cuenta que para obtener los resultados mostrados en la tabla 5, se omitió el paso de estandarizar las variables numéricas. Por ello, la tabla 6 presenta resultados alternativos en un escenario en el que sí se estandarizó las variables numéricas. Los resultados se presentan únicamente para el modelo óptimo (Random Forest entrenado con el conjunto entrenamiento NRO) con los hiperparámetros presentados en el anexo 1.

Tabla 6. Métricas de desempeño del modelo óptimo cuando sí se estandarizaron las variables continuas

|  |  |
| --- | --- |
| **Métrica** | **Random Forest** |
| F1 | 0.654 |
| Accuracy | 0.759 |
| AUC ROC | 0.692 |
| F1 (Sí) | 0.845 |
| F1 (No) | 0.462 |
| 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 NRO). Se muestra que las variables más importantes provienen, mayoritariamente de la base de datos SIAF. Una variable (compu\_muni\_5) proviene de la base de datos RENAMU. No se observa variables políticas ni variables de canon.

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

|  |  |  |
| --- | --- | --- |
| **Variable** | **Importance Score** | **Fuente** |
| \_tejgtotfun\_f5r08pgrcopc | 0.021 | SIAF |
| \_tejgfun\_f5r08ct05pgrcopc | 0.018 | SIAF |
| \_tejgtotfun\_f5pgrco | 0.013 | SIAF |
| \_tejgfun\_f5ct05pgrco | 0.008 | SIAF |
| dfgpimpiafun\_f5ct06pgrco | 0.006 | SIAF |
| dfgpimpiafun\_f5r18ct06pgrco | 0.005 | SIAF |
| \_tejgge\_r09ct05biser | 0.004 | SIAF |
| \_tejgtotfun\_f2opseg | 0.004 | SIAF |
| \_devppimtotfun\_f2transpc | 0.003 | SIAF |
| \_devppimfun\_f1ct05opseg | 0.003 | SIAF |
| \_tejgtotfun\_f5amb | 0.003 | SIAF |
| compu\_muni\_5 | 0.003 | RENAMU |
| devppimtotfun\_f5r07trans | 0.003 | SIAF |
| \_tejgft\_rdet | 0.003 | SIAF |
| devppimtotfun\_f5r07transpc | 0.003 | SIAF |
| \_pimgfun\_f5r18ct06opseg | 0.003 | SIAF |
| \_tejgge\_r08ct05biserpc | 0.003 | SIAF |
| \_tejgfun\_f2ct05opseg | 0.002 | SIAF |
| dfgpimpiage\_r08ct05dotrapc | 0.002 | SIAF |
| Fuente: elaboración propia | | |

El gráfico 1 muestra las correlaciones entre las 20 variables más importantes y las variables de Canon. No se observan correlaciones importantes a excepción de las variables Canon referentes al total ejecutado.

Gráfico 1. Matriz de correlaciones entre las 20 variables más importantes y las variables de Canon

Gráfico

Descripción generada automáticamente

Fuente: elaboración propia

**Mapas de predicción**

Los gráficos 2 y 3 presentan los mapas de predicción según el modelo óptimo (Random Forest entrenado con el conjunto de entrenamiento NRO) para los años 2016 y 2018.

Gráfico 2. Mapas de predicción de acuerdo con el modelo óptimo para el año 2016

Mapa

Descripción generada automáticamente

Fuente: elaboración propia

Gráfico 3. Gráfico 2. Mapas de predicción de acuerdo con el modelo óptimo para el año 2018

Gráfico de dispersión

Descripción generada automáticamente con confianza baja

Fuente: elaboración propia

**ANEXOS**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | **n\_estimators** | **max\_depth** | **max\_features** |
| NRO. Gradient Boosting Trees | 1000 | 2 | 20% |
| NRO. Random Forest | 250 | 30 | 40% |
| Fuente: elaboración propia | | | |

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

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