Dirección de Economía Naranja y Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

El futuro es de todos

DNP
Departamento
Nacional de Planeación

RASTREO DE LAS INVERSIONES PÚBLICAS NACIONALES Y SUBNACIONALES EN GESTIÓN DEL RIESGO DE DESASTRES

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Economía Naranja y Desarrollo Digital - Unidad de Científicos de Datos • Dirección de Ambiente y Desarrollo Sostenible - Subdirección de Cambio Climático y Gestión del Riesgo de Desastres
Sector	Planeación
Tecnologías utilizadas	Python
Fuentes de datos	• Sistema Integral de Información Financiera (SIIF) • Sistema General de Regalías (SGR) • Sistema de Información de Cooperación Internacional (CÍCLOPE)

Contenido

1. Presentación

El seguimiento de las inversiones del orden nacional y subnacional es un punto vital para los diferentes procesos y subprocesos de la gestión del riesgo de desastres, definidos en la Ley 1523 del 2012. Este seguimiento a las inversiones dirigidas a la gestión del riesgo de desastres ayuda a conocer las brechas y oportunidades existentes en la distribución, enfoque y financiamiento. Esta tarea puede llegar a ser tediosa debido a la gran cantidad de proyectos que deben ser revisados y validados año a año por los expertos temáticos. Dada esta necesidad, la Unidad de Cientificos de Datos (UCD) desarrolló una herramienta para realizar rastreos automáticos de proyectos que pueden estar relacionados con la gestión del riesgo de desastres, reportados en tres bases de datos de inversión pública: el Sistema Integrado de Información Financiera (SIIF) Nación, el Sistema General de Regalías (SGR) y la inversión de cooperación internacional a través Cíclope, con una precisión de más del 95%. Esta herramienta también permite consolidar información de indicadores financieros de las bases de datos de inversión pública, reduciendo el trabajo de consolidación de información que debe realizar el experto temático.

The monitoring of national and subnational investments is fundamental for the different processes and sub-processes of disaster risk management defined in Law 1523 of 2012. This monitoring directed to disaster risk management helps to know the existing gaps and opportunities in the distribution, focus, and financing. This task can become tedious due to the large number of projects that must be reviewed and validated year after year by thematic experts. Given this need, the Data Scientists Unit (UCD) developed a tool to perform automatic tracking of projects that may be related to disaster risk management, reported in three public investment databases: the Sistema Integrado de Información Financiera (SIIF) Nación, Sistema General de Regalías (SGR) (SIIF), and international cooperation investment through Cíclope, with a precision of more than 95%. This tool also makes it possible to consolidate information on financial indicators from public investment databases, reducing the work of consolidating information to be carried out by the thematic expert.

2. Objetivos del proyecto

2.1. General

Construir una herramienta de rastreo automático de proyectos de inversión pública en temas relacionados con gestión del riesgo de desastres a través de los nombres de los proyectos que se reportan en las bases de inversión pública del Sistema Integrado de Información Financiera (SIIF) Nación, el Gesproy-Sistema General de Regalías (SGR) y la inversión de cooperación internacional a través Cíclope[1].

2.2. Específicos

1. Implementar algoritmos de limpieza, adecuación y transformación numérica de texto que permitan la identificación de proyectos de inversión pública relacionada con gestión del riesgo de desastres a partir de su título o nombre de proyecto.
2. Implementar un modelo de clasificación supervisado que haga el rastreo automático de proyectos de inversión pública relacionados con gestión del riesgo de desastres.
3. Construir una herramienta que permita realizar el rastreo de nuevos proyectos de inversión pública reportados en las bases SIIF, SGR y CICLOPE; y consolidar información de indicadores de inversión para SIIF y SGR.

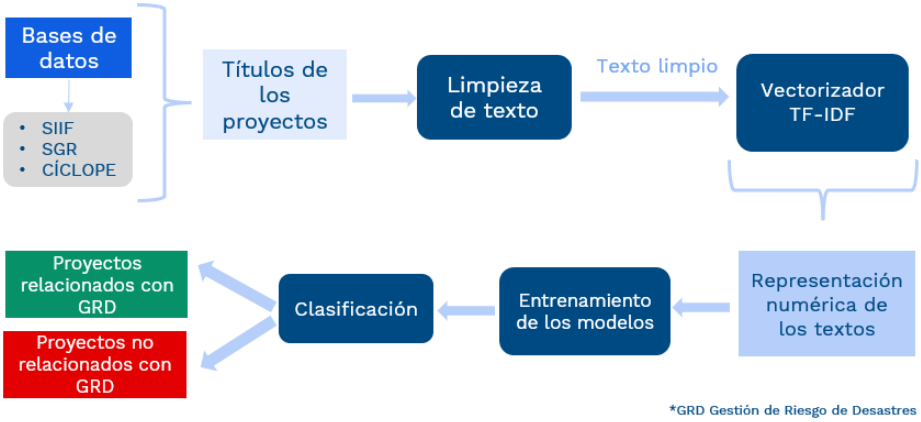
3. Metodología

En esta sección, se detalla la metodología propuesta para el procesamiento de texto de los proyectos reportados en las bases de datos SIIF, Gesproy-SGR y CICLOPE, y la etapa de entrenamiento del rastreador automático de estos proyectos de inversión que están relacionados con gestión del riesgo de desastres, teniendo en cuenta solo el título

[1] Herramienta de consulta que permite visualizar información detallada de los compromisos de programas y proyectos con recursos de Cooperación Internacional No Reembolsable en Colombia, a través de distintos filtros, clasificados según su alcance geográfico: ámbito nacional o departamental y municipal.

del proyecto. Esta metodología se representa en la Figura 1. El primer paso es una limpieza de texto, seguido de una transformación numérica de estos textos limpios. Después, esta transformación numérica se utiliza para entrenar un modelo de rastreo automático. Como etapa final y con el modelo entrenado, se hacen predicciones para nuevos proyectos de inversión provenientes de las tres bases de datos. Estas etapas se explican con mayor detalle a continuación:

Figura 1: Metodología propuesta para el rastreo de proyectos relacionados con gestión del riesgo de desastres.



Fuente: elaboración propia.

3.1. Limpieza de texto

El objetivo de esta etapa es remover todas las palabras que no se consideren útiles para tomar una decisión basada en el texto, por esto, cada título de proyecto reportado en las bases de datos pasa por cinco pasos:

1. Se convierte todo el texto a minúsculas.
2. Se remueven todos los espacios múltiples, saltos de línea, números, signos de puntuación y caracteres especiales tales como @, #, entre otros.
3. Se eliminan las *stopwords* o palabras vacías. Estas hacen referencia a palabras que carecen de sentido cuando se escriben solas. Estas corresponden a conjunciones, artículos, preposiciones y adverbios. Ejemplos de estas palabras son: a, ante, antes, aún, aunque, aquí, arriba, atrás, así, bajo, bastante, cabe, conmigo, bien, casi, cierto, como, de, las, la, el.
4. Quitar palabras propias del contexto: Se remueven del texto palabras que pueden ser muy frecuentes dado la naturaleza del estudio, para ello se removieron nombres propios de personas, municipios y departamentos de Colombia.
5. Se lematiza el texto, esto se refiere a la transformación de todas las formas flexionadas de una palabra en su lema, por ejemplo, canto, cantas, cantamos, cantan son distintas formas (conjugaciones) del verbo cantar, por lo que todas estas palabras son representadas por este lema.

3.2. Vectorización de texto

La vectorización de un texto consiste en generar representaciones vectoriales o numéricas del texto a través de diferentes técnicas, lo cual resulta útil para entrenar modelos de clasificación automática. Para ello se utiliza una metodología de transformación de texto conocida como TF-IDF (*Term Frequency – inverse Document Frequency*), la cual presentó los mejores desempeños en la etapa de rastreo de proyectos. TF-IDF es una medida estadística utilizada para evaluar la importancia de una palabra presente en el nombre un proyecto. Es decir, la importancia de la palabra aumenta proporcionalmente al número de veces que aparece en el título de un proyecto (parte TF), pero se compensa

con la frecuencia de la palabra en los otros títulos que hacen parte del conjunto de entrenamiento (parte IDF). Matemáticamente se representa de la siguiente manera:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Donde $w_{i,j}$ es el peso o importancia de la palabra i en el nombre del proyecto j; $tf_{i,j}$ es el número de ocurrencias de la palabra i en el nombre del proyecto j, df_i es el número de nombres de proyectos en el conjunto de datos de entrenamiento que contienen la palabra i, y N es el total proyectos que hacen parte del conjunto de entrenamiento. En la *Figura 2*, se muestra cómo sería una matriz TF donde el número de filas representa el vocabulario (palabras más frecuentes en los títulos de los proyectos) y en las columnas los títulos de los proyectos. De esta matriz se puede decir que la representación numérica para el título T1 seria $T1 = [1,0,1,0,1, \dots, 0]$, omitiendo la división por el valor de la parte IDF.

Figura 2: Representación a modo de ejemplo de la matriz TF de una vectorización TF-IDF

Títulos en la base de datos

Vocabulario	T1	T2	T3	...	TN
desarrollo	1		3		
sostenible		2	1		1
estudio	1		1		
ambiental			2		1
gestion	1			2	
desastre		1	2		
riesgo	1				2
...					
plan		1		1	

Fuente: Elaboración propia.

La idea detrás de esta representación es encontrar las palabras que mejor modelen la clase positiva (palabras presentes cuando un proyecto está relacionado con gestión del riesgo de desastres) y la clase negativa (palabras presentes cuando un proyecto no es relacionado con gestión del riesgo de desastres).

Después de que cada título (de la clase positiva o negativa) este representado de forma numérica, se procede a una etapa de entrenar un modelo de aprendizaje de máquina que pueda hallar las palabras que mejor describan a cada clase.

3.3. Entrenamiento de modelos de clasificación

En esta etapa se realizan varios experimentos para obtener el modelo que mejor desempeño de clasificación presente a partir de las representaciones numéricas de texto. Para ello se siguieron los siguientes pasos:

1. Probar con diferentes tamaños del vocabulario de la vectorización TF-IDF. En este punto, cada nombre de proyecto se representa por un vector numérico de alta dimensión (como el que se muestra en la *Figura 2*), donde cada dimensión está asociada a un unigrama (palabra) o a un bigrama (dos palabras adyacentes). Las dimensiones se escogen teniendo en cuenta los unigramas y bigramas más frecuentes. Es decir, se ordenan de mayor a menor ocurrencia en los nombres de los proyectos y se asigna una dimensión corte (p. ej, si la dimensión de la representación es 10, esto se traduce en que se escogieron los 10 primeros más frecuentes).

De este paso se obtiene una matriz numérica a partir de los datos de entrenamiento, $X \in \mathbb{R}^{N \times P}$, donde N es el número de nombres de proyectos de la base de datos de entrenamiento y P la dimensión del vocabulario.

2. Con la matriz generada en el paso 1, esta se divide en 2 grupos. Un grupo de entrenamiento (75%) y un grupo de validación (25%). Estos grupos se forman escogiendo de forma aleatoria observaciones dentro de la matriz, garantizando la presencia de ejemplos de nombres de proyectos relacionados con gestión del riesgo de desastres en los dos grupos.
3. Con el grupo de entramiento se ajusta el modelo de clasificación siguiendo una estrategia de validación cruzada de 5 grupos. Esto consiste en dividir este grupo en 5 subgrupos de igual tamaño, se ajusta el modelo con 4 de ellos y se prueba el rendimiento con el restante, esto se hace 5 veces, donde en cada iteración se escoge 1 grupo diferente para la prueba. Esto se hace con el fin de validar, si los parámetros escogidos para el modelo son capaces de generalizar correctamente los datos. Si el rendimiento no es el esperado, se cambian los parámetros del modelo y se repite el procedimiento de validación cruzada. En este proceso se evalúan varias métricas de rendimiento de clasificación, y se escoge la combinación de parámetros que mejor rendimiento promedio haya obtenido.

   Las métricas que se tomaron en cuenta para elegir el modelo de mejor desempeño fueron:
   - ***Precisión***: esta métrica permite conocer el porcentaje de muestras que se han estimado como relacionados con gestión del riesgo de desastres y que realmente lo son (considerando la base de datos de rastreos hechos por el equipo de la Subdirección de Cambio Climático y Gestión del Riesgo de Desastres). El mismo principio se aplica para la clase negativa (o proyectos no relacionados con gestión del riesgo de desastres).
   - ***Recall***: esta métrica es la proporción de los proyectos relacionados con gestión del riesgo de desastres (considerando la base de datos de rastreos hechos por el equipo de la Subdirección de Cambio Climático y Gestión del Riesgo de Desastres) que se estimaron correctamente por el clasificador. El mismo principio se aplica para la clase negativa.
   - **F1-score**: esta métrica combina la *precisión* y el *recall* para obtener un valor más objetivo. Es útil cuando el conjunto de datos no está balanceado, como era el caso de la base de datos de este proyecto.
4. Después se ajusta el modelo con los parámetros de mejor rendimiento sobre todo el grupo de entrenamiento y se experimenta sobre el conjunto de validación, para comprobar que el modelo esté funcionando correctamente.

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación en las bases de datos de entrenamiento y prueba.

- Conjunto de datos extraído del Sistema Integrado de Información Financiera (SIIF) con la información de proyectos relacionados con gestión del riesgo de desastres realizado por el equipo la Subdirección de Cambio Climático y Gestión del Riesgo de Desastres, para los años 2011 hasta el 2021. Esta base de datos cuenta con 362 proyectos relacionados con gestión del riesgo de desastres (clase positiva o etiqueta 1), y 2.760 proyectos no relacionados (clase negativa, esta muestra se toma aleatoriamente de SIIF, tomando en cuenta los proyectos que ya fueron seleccionados por la clase positiva).
- Conjunto de datos extraído del Sistema General de Regalías (SGR) con la información de proyectos relacionados con gestión del riesgo de desastres realizado por el equipo de la Subdirección de Cambio Climático y Gestión del Riesgo de Desastres, para los años 2012 hasta el 2021. Esta base de datos cuenta con 2.086 proyectos relacionados con gestión del riesgo de desastres (clase positiva o etiqueta 1), y 19.010 proyectos no relacionados (clase negativa o etiqueta 0), esta muestra se tomó aleatoriamente de SGR, considerando los proyectos que ya fueron seleccionados por la clase positiva).
- Conjunto de datos extraído de la base de datos CICLOPE con la información de proyectos relacionados con gestión del riesgo de desastres hecho por el equipo de la Subdirección de Cambio Climático y Gestión del

Riesgo de Desastres, para los años 2012 hasta 2018. Esta base de datos cuenta con 130 proyectos relacionados con gestión del riesgo de desastres (clase positiva o etiqueta 1), y 1.265 proyectos no relacionados (clase negativa o etiqueta 0).

4.1. Tamaño del vocabulario

Como se mencionó en la Sección 3.2, se probaron diferentes tamaños de vocabulario hasta encontrar el que mejor describa los proyectos que tienen relación con gestión del riesgo de desastres (clase positiva) y los proyectos que no tienen relación (clase negativa). El tamaño optimo encontrado fue un vocabulario de dimensión 12.000, formado por los unigramas y bigramas más frecuentes en todos los nombres de proyectos registrados en la base de datos (unión de SIIF, SGR CICLOPE).

4.2. Entrenamiento del modelo

En esta etapa se probaron varios modelos de clasificación automática, tales como XGBoost, arboles de decisión, KNN y SVM. El modelo que tuvo mejor rendimiento fue SVM con un kernel RBF, gamma igual a 1.25 y el parámetro de penalización C = 0.15.

Para la etapa de entrenamiento del modelo de rastreo automático de proyectos de inversión, se unieron las tres bases de datos, para un total de 8,813 muestras de títulos de proyectos (relacionados y no relacionados). Estos datos se dividen en dos grupos; el primero el grupo de entrenamiento (75%), con el que se entrena el modelo de rastreo; y el segundo grupo de validación (25%), con el que se valida el correcto funcionamiento y generalización de estimación para nuevos proyectos, evitando el sesgo de entrenamiento. En la Tabla 1, se observa la distribución de los datos por grupo y por clase de pertenencia.

Tabla 1: Distribución de los datos para el entrenamiento del modelo

	Porcentaje de la base de datos	Clase positiva	Clase negativa	Total
Entrenamiento	75%	1.933	17.276	19.209
Validación	25%	645	5.759	6.404

Fuente: elaboración propia.

En la Tabla 2 se observa el rendimiento de estas métricas para el conjunto de prueba (25% del total de los datos), mostrando un buen desempeño tanto en la clase positiva (estimaciones con un 90% de precisión) y la clase negativa (estimaciones con un 99%). Adicional, el acierto promedio del rastreador automático es del 98% sobre el conjunto de prueba, es decir, predijo correctamente 6.276 proyectos (asignándolos como relacionados o no relacionados con gestión del riesgo de desastres) de 6.404.

Tabla 2: Matriz de confusión del modelo

	Precisión	Recall	F1-score	Muestras
Clase negativa	0,99	0,99	0,99	5.759
Clase positiva	0,90	0,93	0,91	645

Fuente: elaboración propia.

5. Herramienta de rastreo automático

El modelo de rastreo automático obtenido en la sección anterior se dispuso en una herramienta que facilita su uso por parte del experto temático. Esta herramienta permite cargar nueva información de proyectos que se reportan en las bases de datos SIIF, Gesproy-SGR y CICLOPE para realizar un rastreo automático a partir de nombre de los proyectos, asignando una probabilidad de pertenencia a temáticas relacionadas con gestión del riesgo de desastres. En la Figura 3, se muestra como se ve la herramienta después de realizar un rastreo automático.

Figura 3: Visualización de resultados de rastreo automático realizado por la herramienta.

Fuente: elaboración propia.

La herramienta también cuenta con un módulo de consolidación de información que el experto temático puede utilizar para obtener cálculos de indicadores financieros sobre las bases de datos (no está disponible para Ciclope).

Figura 4: Modulo de consolidación de información de la herramienta de rastreo.

Fuente: elaboración propia.

6. Conclusiones y recomendaciones

A partir de la metodología desarrollada y de los resultados obtenidos para cumplir los objetivos de este entregable, planteados en el plan de trabajo, se presentan a continuación las principales conclusiones obtenidas por el equipo de la UCD y las principales recomendaciones.

1. El modelo de rastreo automático de proyectos presentó un acierto de clasificación del 98% sobre el conjunto de datos de validación, con una tasa de falsos positivos del 10% y una tasa de falsos negativos del 1%, por lo que la confiabilidad del modelo es alta.
2. Se identificó que proyectos rastreados para la gestión del riesgo de desastres por parte de la Subdirección de Cambio Climático y Gestión del Riesgo de Desastres, no cumplían con la exclusión por filtros negativos dentro de la metodología que se usa actualmente para el rastreo. Es decir, debieron ser rechazados. Por este motivo el clasificador automático mejora el rastreo y reduce la probabilidad de cometer errores debido a rechazo o aceptación por palabras clave.
3. Existen proyectos dentro de los conjuntos de datos de entrenamiento y validación que en su nombre del proyecto no contiene ninguna palabra que tenga relación con gestión del riesgo de desastres, pero que fueron rastreados por el conocimiento del experto temático. Este tipo de observaciones son difíciles de modelar por el rastreador automático por lo que puede ocasionar malas clasificaciones para futuras observaciones con estas condiciones.
4. Se recomienda que siempre se haga una validación los resultados que entrega la herramienta de rastreo, pues, aunque la confiabilidad es alta, existe una probabilidad del 2% de cometer errores de rastreo, especialmente por las condiciones mencionadas en el numeral 3.

7. Socialización

Este proyecto fue socializado con el equipo de la Subdirección de Cambio Climático y Gestión del Riesgo de Desastres, la Dirección de Economia Naranja y Desarrollo Digital y la interior de la UCD.

Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.