Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

Adecuación del algoritmo de rastreo del financiamiento climático realizado por la UCD en 2020, con el fin de utilizarlo para rastrear inversiones del sector público internacional y fortalecimiento del algoritmo para el sector público doméstico para el sistema MRV

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Desarrollo Digital - Unidad de Científicos de Datos • Dirección de Ambiente y Desarrollo Sostenible (DADS)
Sector	Inversión y finanzas públicas
Tecnologías utilizadas	Python – Aprendizaje supervisado – Similitudes entre texto
Fuentes de datos	SGR – SIIF - CICLOPE

Contenido

Presentación

El rastreo de proyectos de inversión relacionados con cambio climático es una tarea de vital importancia debido a la necesidad de generar información precisa para Colombia en el cumplimiento de sus compromisos con el ambiente y el desarrollo sostenible con sus generaciones presentes y futuras, no solo a nivel interno, sino ante la comunidad internacional con la adopción del Acuerdo de París. Por tal motivo, El equipo de la DADS realiza este rastreo de forma semi manual, convirtiéndose en una tarea con un alto gasto de recursos de tiempo y humanos. En pro de alivianar estas cargas, el equipo de la Unidad de Científicos de Datos propuso una metodología automática de rastreo de proyectos relacionados con cambio climático y una herramienta que permite hacer los rastreos de forma rápida, clasificar los proyectos por sector, subsector y destino; además permitir la descarga de los resultados obtenidos mediante el análisis automático.

The tracking of investment projects related to climate change is a task of vital importance due to the need to generate accurate information for Colombia in fulfilling its commitments to the environment and sustainable development with its present and future generations, not only internally, but before the international community with the adoption of the Paris Agreement. For this reason, the DADS team performs this tracking in a semi-manual way, becoming a task with a high expenditure of time and human resources. To alleviate these burdens, the Data Scientist Unit team proposed an automatic tracking methodology for climate change-related projects and a tool that allows quick tracking, classifying projects by sector, sub-sector and destination, and allowing the downloading of the results obtained through automatic analysis.

Objetivos del proyecto

General

Automatizar el proceso de rastreo y categorización de proyectos de inversión relacionados con cambio climático, mediante técnicas de análisis y minería de texto, con el fin de facilitar el accionar de la DADS en esta tarea.

Específicos

1. Explorar las bases de proyectos de inversión del Sistema General de Regalías (SGR) y el Sistema Integrado de Información Financiera (SIIF) previamente categorizadas en una taxonomía sector – subsector – destino por el equipo de la DADS.
2. Evaluar el rendimiento del modelo de clasificación presentado en el año 2020 para bases de datos etiquetadas nuevas tanto de SGR y SIIF como de CICLOPE.
3. Actualizar el algoritmo de aprendizaje supervisado para el rastreo de proyectos de inversión relacionados con cambio climático basado en el título del proyecto, usando técnicas de análisis y minería de texto.
4. Desarrollar una estrategia de similitud de textos entre títulos de proyecto y taxonomía de actividades por sector-subsector -destino y solamente entre títulos de proyectos
5. Desarrollar un tablero de visualización de proyectos rastreados por el algoritmo propuesto.

Metodología

Para realizar el fortalecimiento del algoritmo de clasificación y la adecuación a las nuevas bases de datos se propone una metodología de tres etapas como se ilustra en la *Figura 1*. La primera etapa consiste en evaluar el modelo existente con el fin de verificar si es satisfactorio o no. Para esto, se toman las bases de datos suministradas por la DADS, cuyos proyectos se encuentran previamente identificados como relacionados o no con cambio climático, y se comparan con la clasificación propuesta por el modelo existente. La segunda etapa, es el entrenamiento de un nuevo modelo de aprendizaje automático supervisado, el cual se encarga de seleccionar los proyectos que se relacionan con cambio climático. La tercera etapa, es una estrategia de similitud de textos entre títulos de los proyectos presente en las bases de datos y entre estos y la taxonomía de descripción de actividades de cada sector – subsector – destino.

Adicionalmente, se realiza una estrategia de similitud entre títulos y sus respectivas clasificaciones en sectores, para clasificar a partir de ellos el sector al cual pertenece un proyecto. Los detalles de las tres etapas y la taxonomía se explican a continuación:

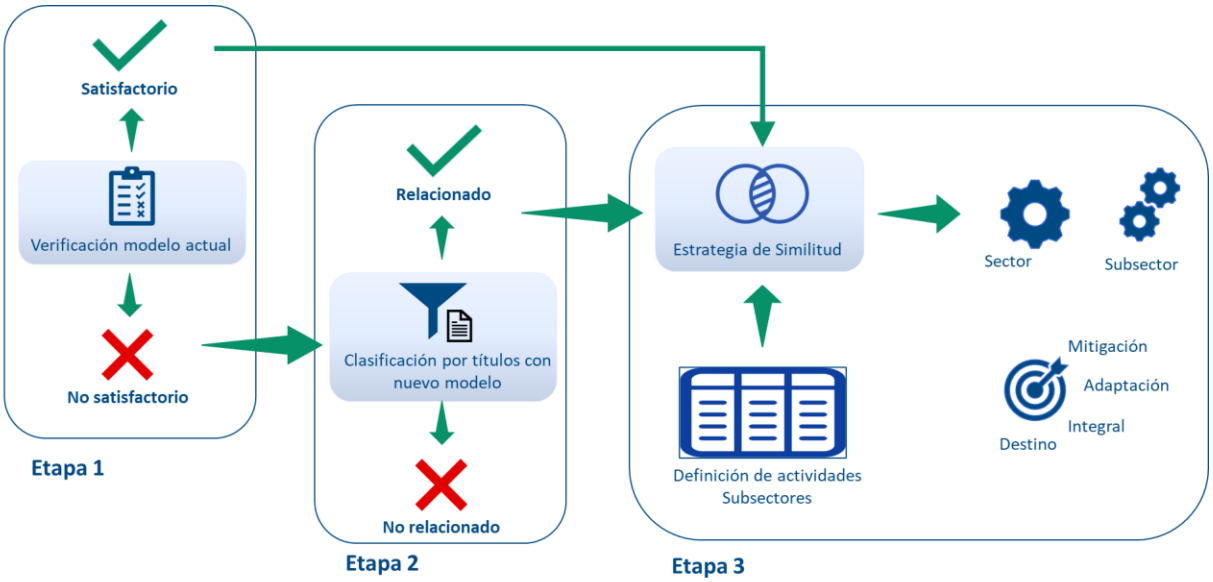


Figura 1 : Metodología en dos etapas para la actualizacion y mejora del modelo de clasificación

Etapa 1: Verificación del modelo anterior

El objetivo de esta etapa es evaluar el modelo de clasificación presentado en el año 2020. Teniendo en cuenta que se suministraron bases de datos de proyectos identificados como relacionados o no con cambio climático, es posible comparar los resultados producto de aplicar el modelo de clasificación a esta información. Los resultados se evalúan con diferentes métricas de rendimiento, en este caso, la métrica que se considera más relevante en el modelo es el *recall*, ésta métrica indica la relación entre los proyectos que el modelo clasifica como relacionados con cambio climático y el total de proyectos clasificados como relacionados (según los proyectos previamente identificados manualmente por el equipo DADS). Para obtener un modelo de clasificación óptimo se busca que el valor de *recall* sea cercano a 1 y el número de proyectos que se rastrean automáticamente como relacionados sea lo menor posible, es decir, que los proyectos clasificados como relacionados sean lo más acertados posible y el número de proyectos clasificados como relacionados, cuando no lo son, sea mínimo.

Etapa 2: Actualización del modelo de clasificación

El objetivo de esta etapa es seleccionar los proyectos que probablemente pueden ser relacionados con cambio climático a partir de su título. Esta tarea inicia con la base de datos que contiene títulos de proyectos relacionados y títulos de proyectos no relacionados, tanto de SGR, SIIF y Cíclope, previamente rastreados y etiquetados por el equipo de la DADS. Cada uno de ellos pasa por un proceso de limpieza de texto, donde se pasa el texto a minúsculas; se hace corrección ortográfica; se eliminan acentos, dobles espacios, signos de puntuación, caracteres especiales, nombres de personas, nombres de departamentos y municipios, números y palabras *stoptwords*[1] del español, con el fin de quitar las palabras que no son relevantes para el análisis de texto y que puedan añadir ruido al proceso de

[1] Stopwords: o palabras vacías, son todas aquellas palabras que carecen de un significado por sí solas. Las stopwords suelen ser artículos, preposiciones, conjunciones, pronombres, entre otros.

entrenamiento del modelo supervisado. Por ejemplo, los nombres de departamentos y nombres de municipios serán palabras frecuentes dada la naturaleza de los proyectos de inversión y no ayudarán en el proceso de rastreo porque están presentes tanto en los proyectos relacionados como en los no relacionados.

Después, cada texto limpio (título después de la limpieza) se transforma a una representación numérica, con el objetivo de aplicar algoritmos de aprendizaje automático. Para ello se utiliza una metodología denominada 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 para un documento en un conjunto de documentos. Es decir, la importancia de la palabra aumenta proporcionalmente al número de veces que aparece en el documento (parte TF), pero se compensa con la frecuencia de la palabra en los otros documentos 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 la importancia de la palabra i en el documento j, $tf_{i,j}$ es el número de ocurrencias de la palabra i en el documento j, df_i es el número de documentos que contienen a i, y N es el total de documentos que hacen parte del conjunto de entrenamiento.

Por último, estas representaciones numéricas de los títulos del proyecto se utilizan para el entramiento del modelo supervisado, que, para el caso específico, fue un SVM (Support Vector Machines) por presentar los mejores desempeños de clasificación. Los resultados obtenidos en esta etapa se consignan en la sección de Resultados.

- ***Taxonomía Sector, Subsector – Destino***

La taxonomía es una tabla de definiciones de las actividades que se realizan en cada sector económico (12 sectores), desagregadas por subsectores (35 subsectores) – destino, que es utilizada como guía para el rastreo semi manual que hace el equipo de la DADS. El destino hace referencia hacía que objetivo está enfocado la inversión, esto puede ser para Mitigación del cambio climático (M), Adaptación (A) o Integral (I). En la *Tabla 1*, se muestra un ejemplo para los sectores Energía y Agropecuario. Estos pueden tener diferentes subsectores y dependiendo de la actividad el destino de la inversión cambia. Este último punto, hace que un mismo subsector pueda tener diferentes tareas, como es el caso del subsector Generación, mejora y acceso a la electricidad, donde tiene una actividad asociada a mitigación y otra a adaptación. En total se tienen 249 actividades distribuidas entre todos los subsectores como se resume en la *Tabla 2*.

Tabla 1: Ejemplo de la taxonomía de descripción de actividades sector-subsector-destino

Sector	Subsector	Actividades	M	A	I
Energía	Generación, mejora y acceso de electricidad	Generar energía con fuentes no convencionales en zonas no interconectadas (sistemas híbridos)	1		
	Generación, mejora y acceso de electricidad	Acceder a la energía a través de la electrificación rural		1	
	Eficiencia energética	Usar energía solar para calentamiento de agua	1		
	Producción minera	Investigar y desarrollar capacidades para mejorar la resiliencia de la actividad minera		1	
Agropecuario	Desarrollo rural	Promover sistemas agroforestales			1

Sector	Subsector	Actividades	M	A	I
	Agricultura	Usar los residuos de cosecha para la generación de energía	1		

Fuente: Elaboración propia

Tabla 2: Distribución de subsectores y actividades por cada sector económico.

ID	Sector	Subsectores	Actividades
1	Agropecuario	4	37
2	Educación	2	3
3	Energía	5	34
4	Gestión del riesgo y atención de desastres	1	8
5	Industria	7	32
6	Medio Ambiente y Recursos Naturales	3	43
7	Residuos	1	15
8	Salud	2	3
9	Transporte	3	40
10	Transversal	3	24
11	Turismo	2	6
12	Vivienda	2	4
	Total	35	249

Fuente: Elaboración propia

Cada actividad de la taxonomía es procesada por la misma rutina de limpieza de texto de la etapa anterior, después cada uno de los textos limpios de las actividades son transformados en vectores numéricos siguiendo la metodología TF-IDF. En este caso, no se tienen en cuenta los títulos de los proyectos de la base de entramiento como en la sección anterior, solo se consideran los textos de las actividades para ajustar el espacio TF-IDF, con dos variaciones:

- Se ajusta el espacio TF-IDF solo teniendo en cuenta las descripciones de las actividades (Método 1).
- Se ajusta el espacio TF-IDF agregando el título de los subsectores a las descripciones de las actividades relacionadas con cada subsector (Método 2).

Al final, se tienen dos representaciones numéricas de la taxonomía, que servirán de insumo en la estrategia de similitud con los títulos de los proyectos de la siguiente etapa.

Etapa 3: Estrategia de similitud título proyecto – taxonomía

Debido a la falta de ejemplos suficientes de proyectos relacionados para cada sector, se propone una estrategia de similitud entre título del proyecto y taxonomía de actividades, con el fin de aproximar futuros proyectos a un sector, subsector y destino.

Partiendo del hecho anterior, cada texto de proyecto limpio es mapeado o transformado a su representación numérica en cada uno de los dos espacios TF-IDF calculados sobre la taxonomía. En este punto, cada representación es un vector que puede ser comparado con los vectores de las actividades de la taxonomía. El supuesto de esta estrategia es que un título de proyecto es más probable de pertenecer al sector – subsector con el que presente mayor similitud entre el vector de la actividad (sea el vector solo de descripción de la actividad o el vector de descripción de la actividad

más título del subsector) y el vector del título. Para cuantificar esas similitudes se utilizan dos medidas de similitud: la similitud coseno para el primer espacio TF-IFD y la similitud de kernel gaussiano para el segundo espacio.

La similitud coseno se define como el producto punto entre los dos vectores numéricos normalizado sobre la multiplicación de las normas de estos:

$$S(i,j) = \frac{v_t^i \cdot v_{tax}^j}{\|v_t^i\| \|v_{tax}^j\|}$$

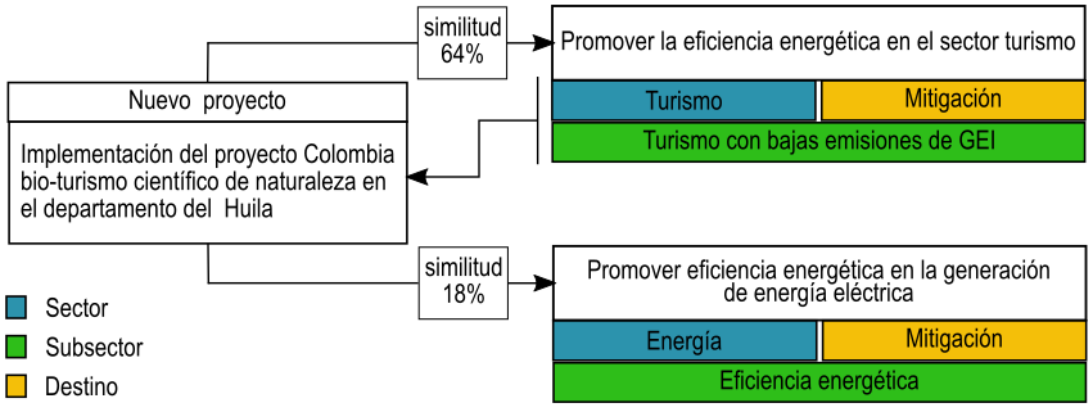
Donde v_t^i es el vector numérico del titulo i, v_{tax}^j es el vector numérico del texto de la actividad j. Por otro lado, la similitud de kernel gaussiano se define como:

$$S(i,j) = \exp(-\gamma \|v_t^i - v_{tax}^j\|^2)$$

La intuición de estas medidas de similitud es: que si dos vectores son exactamente iguales la similitud será 1 y en el caso contrario será 0, lo que representa que no hay ninguna relación entre los vectores.

Al final esta estrategia busca la máxima similitud posible entre título y descripción de la actividad, y el proyecto heredara el sector, subsector y destino de la actividad donde se cumpla esta condición. En la *Figura 2*, se ilustra esta estrategia, donde un nuevo proyecto es comparado con las diferentes actividades de la taxonomía y hereda el sector, subsector y destino de la actividad con la que obtuvo máxima similitud.

Figura 2: Ejemplo de estrategia de similitud. Al nuevo proyecto se lo asignará al sector Turismo, subsector Turismo con bajas emisiones de GEI (Gases de Efecto Invernadero) y con destino de recursos a Mitigación.



Fuente: Elaboración propia.

Adicionalmente, se realizó una metodología que tiene en cuenta solamente el título de los proyectos y la clasificación realizada en ellos previamente. Para ello, se realiza el proceso de vectorización de los títulos de los proyectos generando un espacio vectorial base con el cual se compararán los títulos de los nuevos proyectos. En este caso, para un proyecto nuevo, se comparará la similitud con cada uno de los proyectos del espacio vectorial base y se le asignará el sector de aquel con el que presente mayor similitud.

Resultados

A través del desarrollo metodológico descrito en la Sección 0, se obtuvieron los resultados que se presentan a continuación. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

Etapa 1: Verificación del modelo anterior

- ***Sistemas de información SGR y SIIF***

Las nuevas bases de datos suministradas suman un total de 14486 proyectos rastreados, de los cuales 668 son proyectos clasificados como proyectos relacionados con cambio climático. A partir de las variaciones del límite de decisión (Thresold) se obtiene el número de proyectos que son rastreados y las métricas de rendimiento del modelo para la clasificación de estos proyectos.

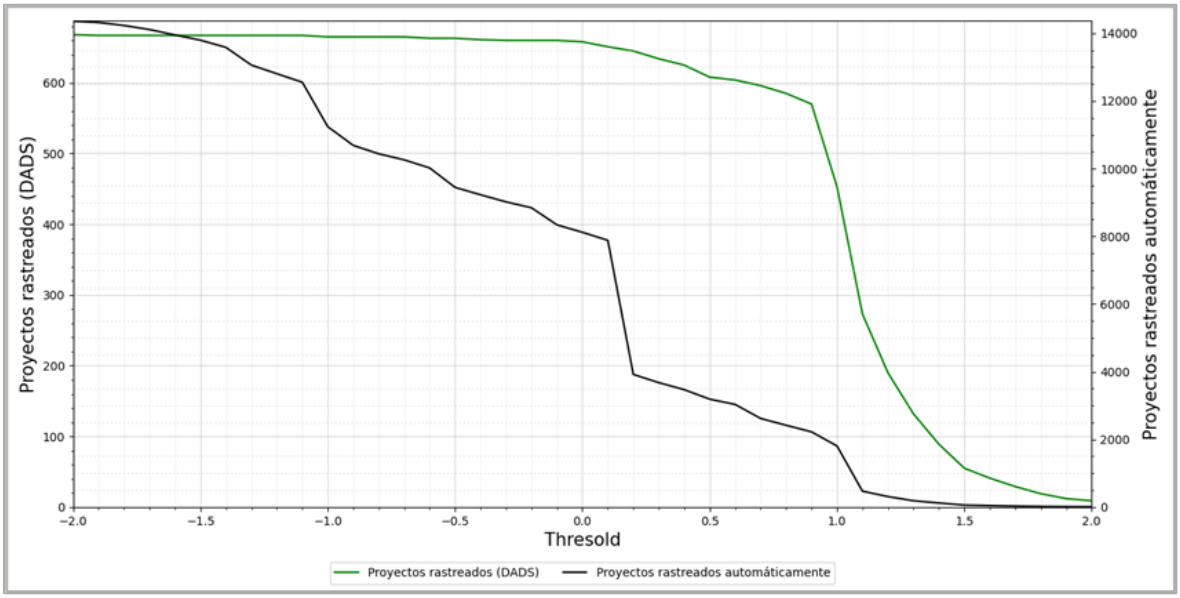


Figura 3: Métricas de rendimiento para base de datos SGR y SIIF del modelo de rastreo anterior.

Como puede observarse en la *Figura 3* el algoritmo de clasificación se comporta satisfactoriamente para valores de *Thresold* entre 0,3 y 0,7. El parámetro principal por evaluar es *recall* (verde en la gráfica), ésta métrica indica la relación entre los proyectos que el modelo rastrea automáticamente como relacionados con cambio climático y el total de proyectos clasificados como relacionados (tomando los proyectos previamente rastreados por la DADS). Según esto, para el intervalo definido anteriormente se pueden trabajar con porcentajes de acierto de clasificación de proyectos relacionados entre el 90% y el 97%. Para estos porcentajes el modelo rastrea entre 2500 a 3500 proyectos respectivamente. Lo anterior indica que con el modelo se puede lograr reducir la cantidad de proyectos por revisar hasta en un 83% (dependiendo del nivel de acierto que la dirección considere aceptable).

- ***Sistemas de información CICLOPE***

La base de datos suministrada suma un total de 924 proyectos rastreados, de los cuales 283 son proyectos clasificados como proyectos relacionados con cambio climático. Siguiendo el procedimiento del inciso anterior se obtienen los siguientes valores de proyectos que son rastreados y sus métricas de rendimiento asociadas:

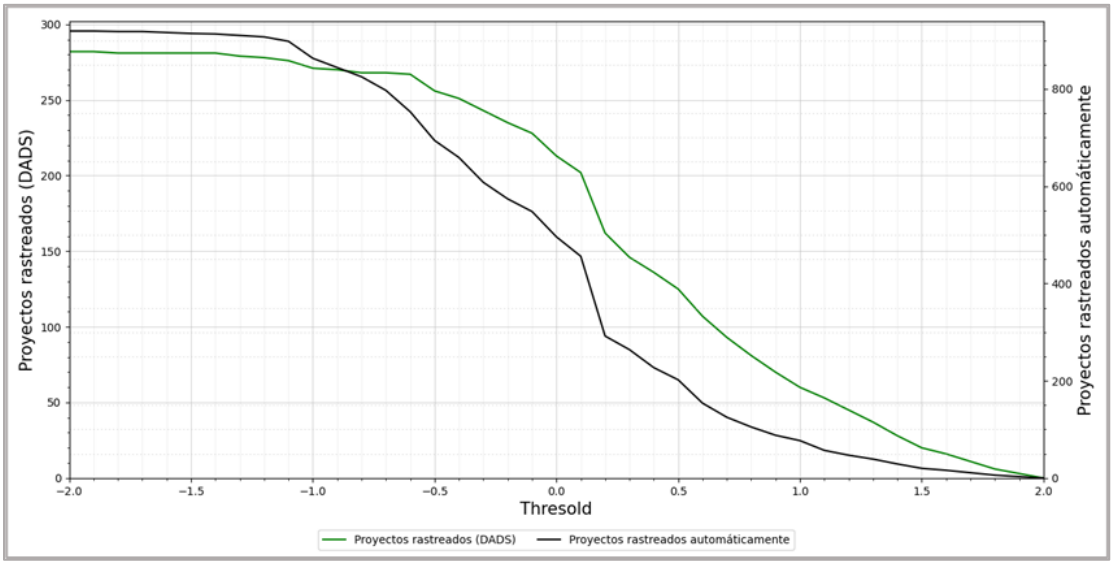


Figura 4: Métricas de rendimiento base de datos cíclope con el modelo de rastreo actual.

Como se ve en la *Figura 4* el algoritmo de clasificación no se adapta de manera adecuada a la base de datos Cíclope. Para obtener rendimientos mayores al 90% es necesario valores de Thresold por debajo de -0.5 lo que implicaría un rastreo de 694 proyectos (una reducción de sólo el 25%).

Teniendo en cuenta los resultados anteriores, se tomó la decisión de realizar un modelo de rastreo automático solo para esta base de datos, con el objetivo de mejorar el proceso de rastreo.

Actualización del modelo de rastreo automático para SGR y SIIF

Para entrenar el modelo para los sistemas de información SGR y SIIF se contó con una base de proyectos previamente identificados como relacionados, 706 títulos de proyectos de SRG y 941 títulos de proyectos SIIF, para un total de 1647 proyectos. Para la clase negativa, se cuenta con 20619 proyectos identificados como no relacionados de las bases de datos SGR y SIIF, completando 22266 proyectos. A partir de estos proyectos base se calcula el espacio de transformación numérica TFiDF con 550 dimensiones. Por último, se entrenó un modelo de clasificación binario (1- relacionado, 0 – no relacionado) SVM (Support Vector Machines) en una metodología de validación cruzada con 5 folds con el 75% de los proyectos base (16699), el 25% (5567) restante se dejó para validación del rendimiento del clasificador. En la *Tabla 3*, se muestran los resultados del modelo para los 5567 proyectos.

Tabla 3: Métricas de evaluación del modelo de rastreo de proyectos relacionados con cambio climático (SGR -SIIF).

	Precision	Recall	F1-score	Muestras
No relacionado	0.99	0.99	0.99	5160
Relacionado	0.86	0.86	0.86	407
Promedio ponderado	0.98	0.98	0.98	5567

Fuente: Elaboración propia.

Al tratarse de una clasificación con una muestra desbalanceada de clases es normal presentar valores algo más bajos en las métricas de evaluación del modelo (*Tabla 3*) para la clase no predominante (relacionados con cambio climático). Sin embargo, se puede apreciar que el proceso de clasificación como relacionado/no relacionado tiene un acierto del

98% sobre el 25% de los proyectos de la base de datos, adicionalmente, como se ve en la *Figura 5* el modelo presenta un buen comportamiento comparado con el que se contaba anteriormente (*Figura 3*).

Como se observa en la *Figura 5* el algoritmo de clasificación presenta mejoras considerables y un comportamiento óptimo para valores de *thresold* entre -0,9 y 0,9. Según esto, para el intervalo definido anteriormente, el algoritmo de rastreo automático identifica de 716 a 771 proyectos relacionados con inversión en cambio climático, en los cuales se encuentran del 97% al 98% de los proyectos identificados como relacionados por la DADS. Lo anterior indica que con el modelo se puede lograr una reducción en la cantidad de proyectos por revisar hasta en un 97% (dependiendo del nivel de acierto que la dirección considere aceptable). Teniendo en cuenta la *Figura 5* un valor recomendable para el *thresold* es de 0.

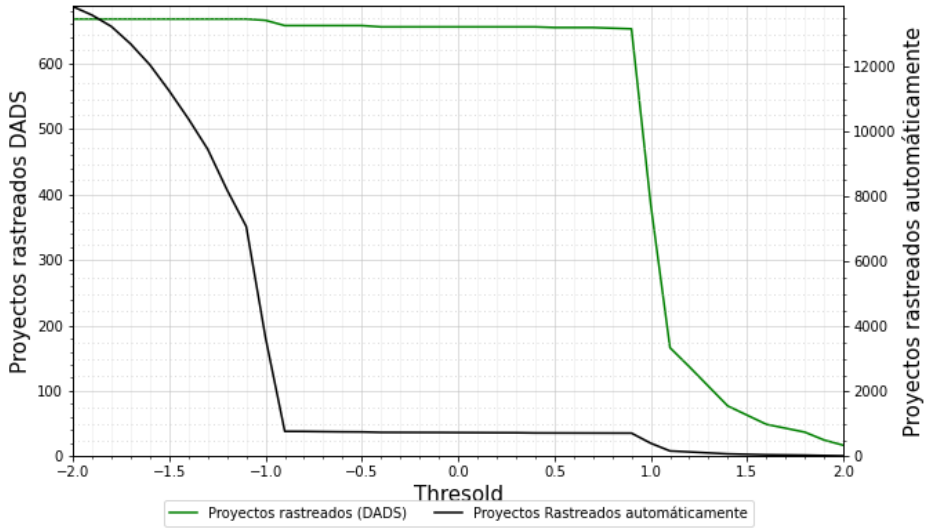


Figura 5: Desempeño del algoritmo sobre el número de proyectos de la base de datos SGR-SIIF

Por otro lado, para el sistema de información CICLOPE se cuenta con una base de datos con 283 proyectos clasificados como relacionados y un restante de 641 proyectos no relacionados con cambio climático, para un total de 924 proyectos. A partir de estos proyectos base se calcula el espacio de transformación numérica TFiDF con 330 dimensiones. Por último, se entrenó un modelo de clasificación binario (1- relacionado, 0 – no relacionado) SVM (Support Vector Machines) en una metodología de validación cruzada con 5 folds con el 80% de los proyectos base (785), el 20% (139) restante se dejó para validación del rendimiento del clasificador. En la *Tabla 3*, se muestran los resultados del modelo para los 139 proyectos.

Tabla 4: Métricas de evaluación del modelo de rastreo de proyectos relacionados con cambio climático (SGR -SIIF).

	Precision	Recall	F1-score	Muestras
No relacionado	0.91	0.90	0.90	96
Relacionado	0.77	0.79	0.78	43
Promedio ponderado	0.86	0.86	0.86	139

Fuente: Elaboración propia.

De manera similar al análisis que se realizó para los sistemas de información SGR y SIIF, si bien las métricas de desempeño del proyecto no se muestran muy altas para la clase no predominante (relacionados con cambio climático),

el rendimiento general de clasificación como relacionado/no relacionado es bueno y presenta un acierto del 86% sobre el 20% de la base de datos de entrenamiento. Esto se evidencia en la *Figura 6* al compararla con la *Figura 4*.

En este caso, se identifica un comportamiento aceptable del modelo para valores de thresold entre -0.2 y 0. Pare este intervalo el modelo rastrea entre 340 y 287 proyectos respectivamente que están relacionados con cambio climático, logrando aciertos entre 92% y 88% y reduciendo la cantidad de proyectos que deben ser revisados hasta en un 69%.

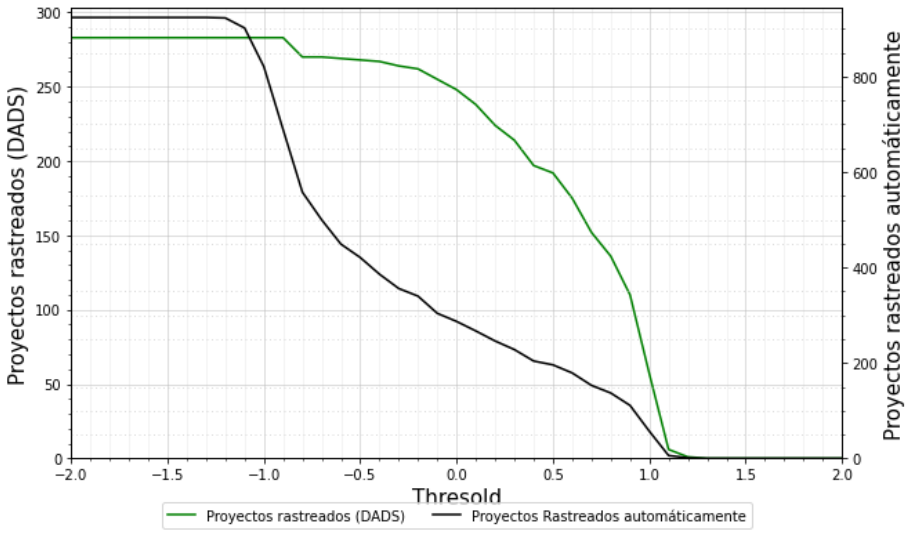


Figura 6: Desempeño del algoritmo sobre el número de proyectos de la base de datos CICLOPE

Estrategias de similitud

Método 1: similitud solo considerando descripciones de las actividades (M1):
La estrategia de similitud considerando las descripciones de las actividades no presenta mejoras considerables al actualizar la base de datos, razón por la cual se mantienen el mismo modelo presentado en el 2020. Para esta parte se obtuvo un acierto de 51% de 979 archivos relacionados por sectores. Se ajustó un modelo Bag of Words de dimensión 1735. La *Tabla 5* resume los rendimientos por sector. Los sectores de industria, salud, turismo y vivienda no se pudieron cuantificar de forma correcta debido a la falta de proyectos ejemplo, no obstante, esta estrategia permite clasificar en el futuro proyectos relacionados con estos sectores.

Tabla 5: resumen de rendimiento de categorización mediante la estrategia de similitud título vs descripción de subsectores (SGR - SIIF)

Sectores	Precision	Recall	F1 - score	Muestras
Agropecuario	0.51	0.48	0.49	98
Educación	0.17	0.23	0.19	13
Energía	0.26	0.49	0.34	84
Gestión del riesgo y atención de desastres	0.75	0.30	0.43	132
Industria	0.04	0.20	0.07	5
Medio Ambiente y Recursos Naturales	0.64	0.67	0.66	422
Residuos	0.71	0.26	0.38	78
Salud	0.00	0.00	0.00	0
Transporte	0.74	0.89	0.01	54

Sectores	Precision	Recall	F1 - score	Muestras
Transversal	0.38	0.22	0.28	92
Turismo	0.00	0.00	0.00	1
Vivienda	0.00	0.00	0.00	0
Promedio	0.59	0.51	0.52	979

Fuente: Elaboración propia.

Para el caso de la base de datos CICLOPE se obtiene un acierto del 45% de 283 proyectos relacionados por sectores. Se ajustó un modelo Bag of Words de dimensión 1860. La *Tabla 6* resume los rendimientos por cada sector. Los sectores de Educación, Salud y Turismo no pudieron ser cuantificados de forma adecuada debido a la falta de proyectos de ejemplo clasificados en estos sectores. Sin embargo, en el futuro puede actualizarse estos insumos para mejorar el rendimiento del modelo en estos sectores.

Tabla 6: resumen de rendimiento de categorización mediante la estrategia de similitud título vs descripción de subsectores (CICLOPE)

Sectores	Precision	Recall	F1 - score	Muestras
Agropecuario	0.43	0.39	0.41	38
Educación	0.00	0.00	0.00	6
Energía	0.64	0.35	0.45	26
Gestión del riesgo y atención de desastres	0.60	0.33	0.43	18
Industria	0.19	0.75	0.30	4
Medio Ambiente y Recursos Naturales	0.56	0.61	0.58	84
Residuos	0.50	0.75	0.60	4
Salud	0.00	0.00	0.00	0
Transporte	0.39	0.80	0.52	15
Transversal	0.42	0.32	0.36	84
Turismo	0.08	1.00	0.15	1
Vivienda	0.25	0.33	0.29	3
Promedio	0.48	0.45	0.45	283

Fuente: Elaboración propia.

Método 2: similitud considerando nombre subsector + descripción (M2):

De la misma manera, el rendimiento se mantiene más alto para el modelo existente y este se conserva. Esta estrategia presentó resultados similares obtenidos por el modelo de similitud M1, 52% de acierto sobre los 979 proyectos relacionados por sector bajo un modelo TFiDF de dimensión 2005. Puede presentar mejores resultados de estimación de sectores debido a que cuenta con más palabras relacionadas provenientes de los títulos de los subsectores, pero las medidas de similitud son más bajas por el mismo hecho. La *Tabla 7* muestra los resultados obtenidos por esta segunda metodología.

Tabla 7: resumen de rendimiento de categorización mediante la estrategia de similitud título – descripción de subsectores + título subsectores (SGR-SIIF)

Sectores	Precision	Recall	F1 - score	Muestras
Agropecuario	0.51	0.52	0.52	98

Sectores	Precision	Recall	F1 - score	Muestras
Educación	0.38	0.46	0.41	13
Energía	0.36	0.36	0.36	84
Gestión del riesgo y atención de desastres	0.75	0.38	0.50	132
Industria	0.00	0.00	0.00	5
Medio Ambiente y Recursos Naturales	0.69	0.68	0.69	422
Residuos	0.61	0.26	0.36	78
Salud	0.00	0.00	0.00	0
Transporte	0.69	0.81	0.75	54
Transversal	0.38	0.26	0.31	92
Turismo	0.00	0.00	0.00	1
Vivienda	0.00	0.00	0.00	0
Promedio	0.61	0.52	0.55	979

Fuente: Elaboración propia.

Para la base de datos CICLOPE el modelo presenta un acierto del 45%. En la *Tabla 8* se muestra el resumen de los rendimientos para cada sector para un modelo TFiDF de dimensión 2120. Como se puede notar, se encuentran los mismos inconvenientes para los sectores con baja información.

Tabla 8: Resumen de rendimiento de categorización mediante la estrategia de similitud título – descripción de subsectores + título subsectores (CICLOPE)

Sectores	Precision	Recall	F1 - score	Muestras
Agropecuario	0.46	0.34	0.39	38
Educación	0.00	0.00	0.00	6
Energía	0.67	0.54	0.60	26
Gestión del riesgo y atención de desastres	0.57	0.44	0.50	18
Industria	0.23	0.75	0.35	4
Medio Ambiente y Recursos Naturales	0.64	0.63	0.63	84
Residuos	0.14	0.25	0.18	4
Salud	0.00	0.00	0.00	0
Transporte	0.33	0.80	0.47	15
Transversal	0.47	0.35	0.40	84
Turismo	0.17	1.00	0.29	1
Vivienda	0.12	0.33	0.18	3
Promedio	0.51	0.48	0.48	283

Fuente: Elaboración propia.

Método 3: similitud considerando el título de los proyectos

Este método tuvo en cuenta la creación de un espacio vectorial de dimensión 2225 con un modelo TFiDF y con un acierto del 84%. Correspondiente a 1316 proyectos que funcionaron como entrenamiento del modelo (80% del total de proyectos relacionados y clasificados por sector y subsector). Con el 20% de los proyectos restantes (329) se realizó la prueba del modelo de clasificación obteniendo los rendimientos que se presentan en la *Tabla 9*. Allí se destacan

algunos sectores donde la falta de proyectos clasificados en ellos no permite clasificarlos en estos sectores (Salud, Turismo y Vivienda) y donde la baja información en ellos puede presentar resultados no muy confiables (Industria).

Tabla 9: Resumen de rendimiento de categorización mediante la estrategia de similitud de títulos (SGR-SIIF)

Sectores	Precision	Recall	F1 - score	Muestras
Agropecuario	0.77	0.75	0.76	32
Educación	0.60	0.75	0.67	4
Energía	0.93	0.90	0.92	31
Gestión del riesgo y atención de desastres	0.89	0.96	0.93	71
Industria	0.00	0.00	0.00	1
Medio Ambiente y Recursos Naturales	0.86	0.89	0.88	118
Residuos	0.58	0.58	0.58	19
Salud	0.00	0.00	0.00	0
Transporte	0.81	0.94	0.87	18
Transversal	0.92	0.68	0.78	34
Turismo	0.00	0.00	0.00	1
Vivienda	1.00	1.00	1.00	1
Promedio	0.84	0.84	0.84	330

Fuente: Elaboración propia.

Para el caso de CICLOPE se presenta un espacio de dimensión 1800 con un modelo Bag of Words y con un acierto del 51%. Los rendimientos para cada sector se pueden ver en la *Tabla 10*. Allí se identifica que los sectores de Educación, Industria, Residuos, Salud, Transporte y Vivienda no cuentan con una base de datos amplia que le permita al modelo desempeñarse de mejor manera y entrenarse mejor.

Tabla 10: Resumen de rendimiento de categorización mediante la estrategia de similitud de títulos (CICLOPE)

Sectores	Precision	Recall	F1 - score	Muestras
Agropecuario	0.50	0.38	0.43	8
Educación	0.00	0.00	0.00	1
Energía	0.67	0.40	0.50	5
Gestión del riesgo y atención de desastres	0.50	0.50	0.50	4
Industria	0.00	0.00	0.00	1
Medio Ambiente y Recursos Naturales	0.65	0.76	0.70	17
Residuos	0.00	0.00	0.00	1
Salud	0.00	0.00	0.00	0.00
Transporte	0.00	0.00	0.00	0.00
Transversal	0.67	0.67	0.67	3
Turismo	0.41	0.41	0.41	17
Vivienda	0.00	0.00	0.00	0.00
Promedio	0.52	0.51	0.51	57

Fuente: Elaboración propia.

Herramienta de visualización

Por último, se desarrolló una herramienta de visualización que permite cargar una base de datos de proyectos, ya sea SGR, SIIF o CICLOPE, seleccionar el modelo de clasificación con el que se desea trabajar, seleccionar la metodología de similitud, mostrar los proyectos que han sido rastreados y descargar los resultados con base en las elecciones del usuario de acuerdo a los sectores de interés y el límite de decisión sobre cuándo un proyecto nuevo es asignado como relacionado con cambio climático, este valor está predeterminado para cada modelo de clasificación (optimo sobre la base de datos de entrenamiento), pero permite que el usuario elija el balance entre falsos positivos (valores a la izquierda del valor por defecto) y falsos negativos (valores a la derecha del valor por defecto). Entre más a la derecha se encuentre el cursor de límite de decisión seleccionado por el usuario, el modelo tendrá más precisión en seleccionar los proyectos que son relacionados con cambio climático. En el caso contrario, si el puntero se encuentra hacia la izquierda, existirá mayor probabilidad de categorizar un proyecto como relacionado cuando en realidad no lo es. En la Figura 7 se muestra una captura de la herramienta.

Figura 7: Captura de pantalla de la herramienta de visualización

Conclusiones y recomendaciones

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

1. La constante actualización y ampliación de la base de datos puede mejorar cada vez más el rendimiento del modelo de clasificación, permitiendo mejorar los resultados de este y facilitando el proceso de rastreo de proyectos relacionados con cambio climático. Prueba de ello son los rendimientos del nuevo modelo para SGR-SIIF que muestra mejores rendimientos para bases de datos más grandes.
2. La tasa de acierto del modelo supervisado puede estar sesgado a términos frecuentes presentes en los documentos de entrenamiento. Para lograr una mejor confiabilidad se aconseja lograr un balance entre la tasa de falsos positivos y falsos negativos aceptados. Además de seguir alimentando la base de datos sobre los sectores y subsectores que no se encontraron presentes en la base de entrenamiento.

3. Para lograr mejorar el desempeño de la estrategia de similitud, tanto en la metodología 1 y 2, se aconseja alimentar las descripciones de las taxonomías con más términos relacionados a cada subsector y eliminar las actividades redundantes inter e intra-sectores.
4. Para lograr mejorar el desempeño de la estrategia de similitud de la metodología 3, se aconseja actualizar constantemente la base de datos que alimenta el modelo de similitud. Ampliando de esta forma los términos comunes que aparecen en los títulos de los proyectos en cada sector.
5. La herramienta de visualización permite a los usuarios de la DADS seleccionar el límite de decisión que más se ajuste dependiendo de la base de datos SGR, SIIF o CICLOPE.
6. Como mejora futura, se puede conseguir información adicional a los títulos de los proyectos provenientes de bases de datos de una actualización frecuente, debido a la falta de información en bases de datos como Mapa de Inversión. Esto con el fin de mejorar los modelos de clasificación y de similitud.

Socialización

Los resultados del presente proyecto se socializaron con la Dirección de Ambiente y Desarrollo Sostenible DADS, la Unidad de Científicos de Datos, el Director de Desarrollo Digital y el Subdirector General Sectorial.

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.