Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

ACTUALIZACIÓN Y FORTALECIMIENTO DEL ALGORITMO DE RASTREO DEL FINANCIAMIENTO CLIMÁTICO (ETAPA 3)

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
Sector	Inversión y finanzas públicas
Tecnologías utilizadas	Python – Aprendizaje supervisado – Similitudes coseno
Fuentes de datos	SGR – SIIF - CÍCLOPE

Contenido

1. Presentación 2
2. Objetivos del proyecto 2
3. Metodología 2
4. Resultados 7
5. Conclusiones y recomendaciones 10
6. Socialización 11
Contacto 11

1. 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 del 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, clasificación en dimensiones de adaptación, y permitir la clasificación por líneas estratégicas e instrumentales de la Política Nacional de Cambio Climático (PNCC); 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, classification in adaption dimensions, and allow the classification by strategic and instrumental lines of the National Policy for Climate Change (NPCC); also allow the downloading of the results obtained through automatic analysis.

2. Objetivos del proyecto

2.1. General

Fortalecer la clasificación de proyectos o iniciativas climáticas como insumo básico para el sistema MRV de financiamiento climático mediante la actualización del algoritmo de rastreo de proyectos ya existente e inclusión de nuevas funcionalidades para la inclusión de asociación de líneas estratégicas e instrumentales y dimensiones de adaptación.

2.2. Específicos

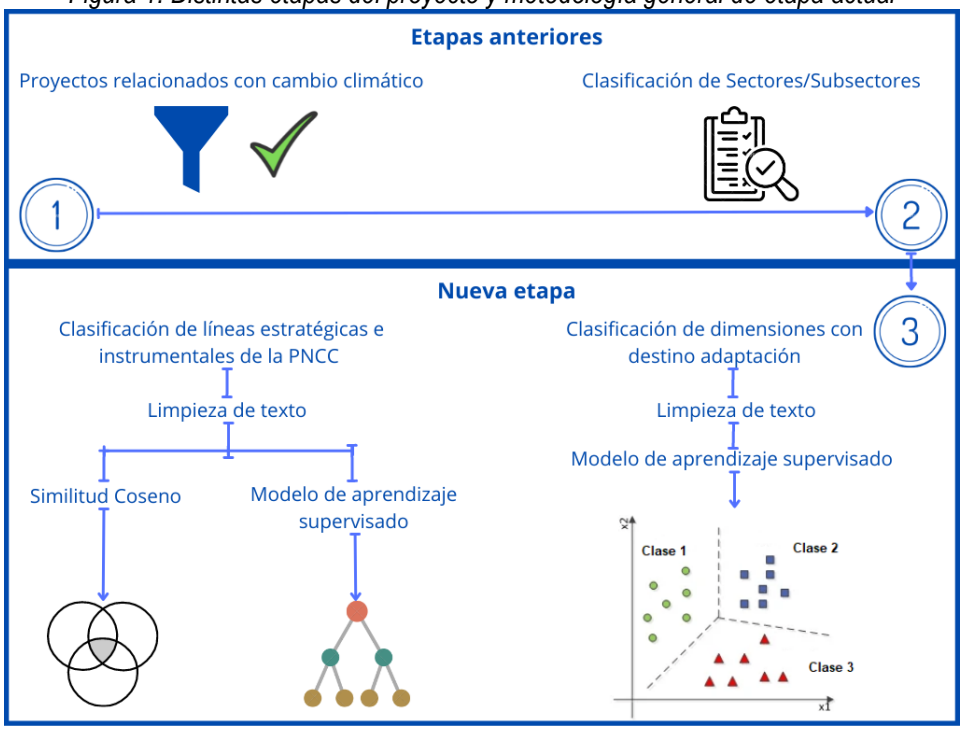
1. Validar el rendimiento de los modelos de asignación de sectores, subsectores y destino del algoritmo actual.
2. Ajustar modelos de asignación de sectores, subsectores y destino para mejorar su rendimiento para las fuentes de proyectos SIIF, SGR y Cíclope.
3. Implementar algoritmo automático de asociación de las líneas estratégicas de la Política Nacional de Cambio Climático (PNCC) a partir del título de los proyectos y las acciones de las líneas mediante técnicas de analítica de texto.
4. Asociar las dimensiones con destino adaptación a través de los títulos o nombres de las actividades mediante técnicas de analítica de texto.
5. Actualizar la herramienta de rastreo de proyectos de cambio climático.

3. Metodología

Para realizar la inclusión de nuevas funcionalidades para el algoritmo de clasificación se propone una metodología ilustrada en la Figura 1 en donde se presentan diferentes etapas de la totalidad del proyecto hasta ahora. En las anteriores etapas del proyecto se procedió de manera que primero se observaban los proyectos relacionados al cambio climático por medio de un modelo de aprendizaje supervisado, luego con la clasificación de sectores/subsectores teniendo en cuenta la similitud de proyectos ya clasificados y las descripciones de sectores/subsectores, ahora en la presente etapa se procedió a clasificar los proyectos en líneas estratégicas e instrumentales, para esto se probaron dos metodologías, la generación de un modelo de aprendizaje supervisado y por similitud de coseno; finalmente para

la parte de clasificación de dimensiones de adaptación se usó solamente la metodología de modelo de aprendizaje supervisado.

Figura 1: Distintas etapas del proyecto y metodología general de etapa actual



Fuente: Elaboración propia

A continuación, se presentan la cantidad de datos según clasificación de dimensiones en la Tabla 3 y líneas estratégicas e instrumentales en la Tabla 1 y Tabla 2 respectivamente, para cada sistema de información y que se usaron en las metodologías que se van a comentar a continuación.

Tabla 1: Cantidad proyectos en líneas estratégicas

Línea	Nombre	Ciclope	SGR	SIIF
LE_1	Desarrollo rural bajo en carbono y resiliente al clima	69	84	63
LE_2	Desarrollo urbano bajo en carbono y resiliente al clima	66	88	25
LE_3	Desarrollo minero-energético bajo en carbono y resiliente al clima	0	18	17
LE_4	Desarrollo de infraestructura baja en carbono y resiliente al clima	41	1	19
LE_5	Manejo y conservación de ecosistemas y SE para un desarrollo bajo en carbono y resiliente al clima	216	72	170

Fuente: Elaboración propia

Tabla 2: Cantidad proyectos en líneas instrumentales

Línea	Nombre	Cíclope	SGR	SIIF
LI_1	Planificación de la gestión del cambio climático	61	13	91
LI_2	Información y ciencia, tecnología e innovación	68	23	30
LI_3	Educación, formación y sensibilización a públicos	6	7	7
LI_4	Financiación e instrumentos económicos	0	1	8

Fuente: Elaboración propia

Tabla 3: Cantidad proyectos en dimensiones de adaptación

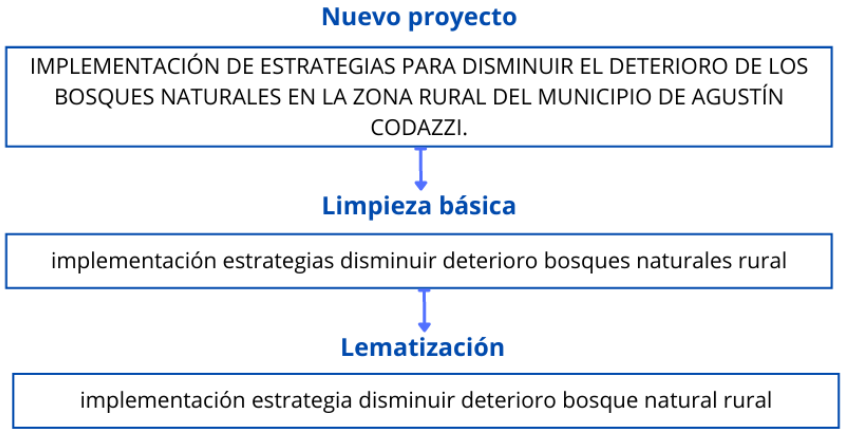
Dimensión	Nombre	Cíclope	SGR	SIIF
D_1	Biodiversidad y servicios ecosistémicos	103	121	229
D_2	Hábitat humano	89	81	134
D_3	Infraestructura	19	121	47
D_4	Recurso hídrico	37	276	223
D_5	Seguridad alimentaria	31	53	6
D_6	Salud	2	0	0

Fuente: Elaboración propia

3.1. Limpieza de texto

La limpieza de texto se usó para los títulos de los proyectos; títulos, subtítulos, acciones de las líneas en todos los sistemas de información para la generación de los modelos y en el algoritmo para la adecuada clasificación de nuevos proyectos.

Figura 2: Ejemplo de limpieza de texto



Fuente: Elaboración propia

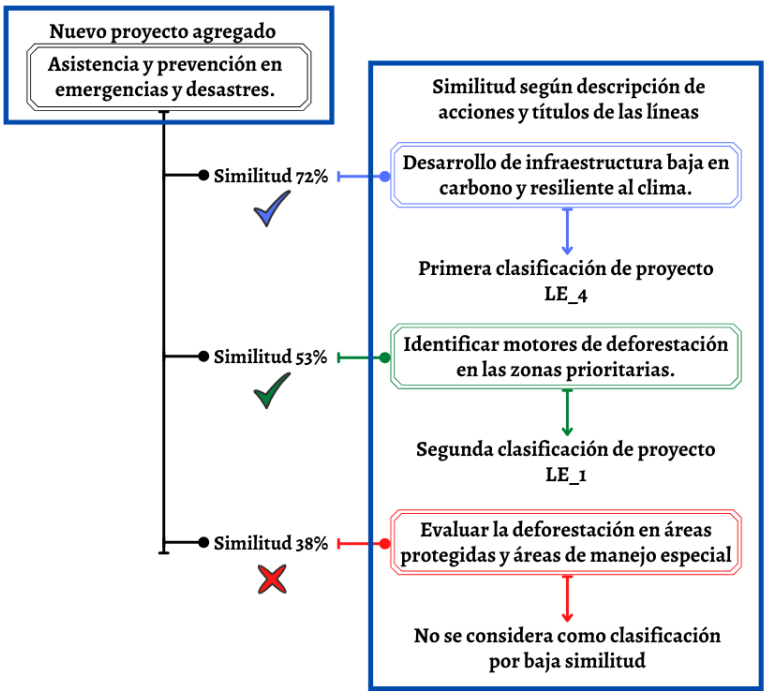
El proceso se puede ver como ejemplo en la Figura 2, en el que primero que todo se implementa una limpieza básica que vuelve en minúscula el texto, se remueven los signos de puntuación y palabras stopwords, las cuales son palabras que consideramos que pueden afectar negativamente la interpretación del texto por parte del algoritmo, a partir de esa limpieza básica se inicia el proceso de lematización que puede convertir esas palabras derivadas de otras en su raíz, ósea devuelve la forma que tienen esas palabras cuando las buscamos en un diccionario. La lematización puede llegar a ser un proceso clave en el modelo de clasificación, pero tiene dos costos. Primero, es un proceso que consume recursos (sobre todo tiempo). Segundo, suele ser probabilística, así que en algunos casos podemos obtener resultados inesperados.

3.2. Clasificación con metodología similitud coseno

Esta metodología fue usada solamente para las líneas estratégicas e instrumentales, ya que al notar que el rendimiento no era tan bueno como el del modelo de aprendizaje supervisado se decide solo tener en cuenta este último, entonces para la metodología de similitud coseno se usaron las descripciones de acciones, líneas de acción y el mismo título de la línea para las líneas estratégicas y en el caso de líneas instrumentales se usó los subtítulos PNCC con los títulos de las líneas, respectivamente se le hizo una concatenación a cada acción, título o subtítulo por medio de una lista a cada línea. A estas se le aplicó la limpieza de texto mencionada en el apartado 3.1.

Después se pasó el texto a formato numérico por medio del vectorizador Term frequency – Inverse document frequency (Tf-idf), que a su vez procederá a la estrategia de similitud en donde para comprobar su funcionamiento correcto se ingresaron los títulos ya clasificados suministrados por la DADS, con un total de 949 proyectos para líneas estratégicas y un total de 315 proyectos para líneas instrumentales.

Figura 3: Ejemplo de clasificación en líneas estratégicas por medio de la estrategia de similitud



Fuente: Elaboración propia

A fin de observar el rendimiento de esta estrategia de similitud coseno que permite comparar la representación numérica o vector de un proyecto que se desea clasificar con las otras representaciones vectoriales de las descripciones de acción y títulos, esta comparación vectorial se hace por medio del ángulo entre ellos, a ese ángulo le sacamos el coseno que nos devuelve un valor entre el rango de 0 a 1, siendo 1 una similitud perfecta y 0 que sencillamente no hay ningún documento a la que se parezca, se usarán los dos valores más altos en la clasificación de las líneas, para comprobar si se está dando la clasificación de manera adecuada se hace una comparación entre la base de datos con los proyectos adecuadamente clasificados y los títulos que acabamos de clasificar por medio de similitud. Este proceso se puede observar de manera gráfica en la Figura 3 en donde la similitud es representada como un porcentaje para comprender con mayor facilidad.

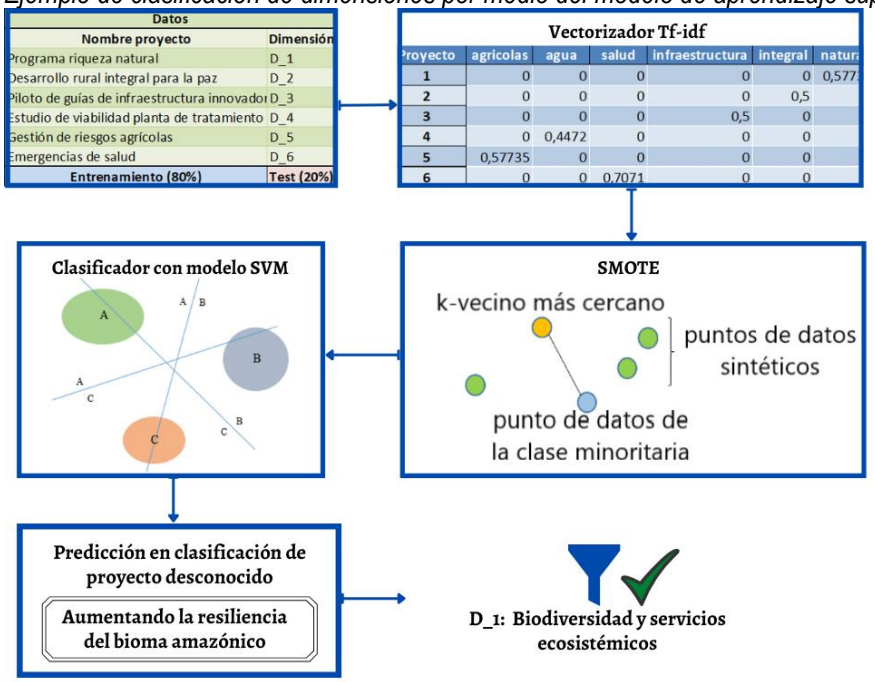
En este caso se variaron los parámetros referentes a los n-gramas y la cantidad máxima de caracteres que debe agarrar el vectorizador, estas variaciones se hicieron a partir de varias iteraciones y observando los resultados que se producían a partir de matrices de confusión, y otras métricas como precision, recall, F1-Score y accuracy.

3.3. Clasificación con metodología por modelo de aprendizaje supervisado

Dado que la similitud coseno llega a fallar por la poca información que se le estaba suministrando, especialmente en el caso de líneas instrumentales que solo estaban recibiendo su propio título y un subtítulo se decide priorizar esta metodología para el caso de las líneas y dimensiones, en la cual se procedió con los datos suministrados de títulos ya clasificados implementado su limpieza de texto, después agarramos el 80% de esos datos para el entrenamiento del modelo, el 20% restante lo guardamos para testear o comprobar el rendimiento del modelo, después se pasaron los textos de la parte de entrenamiento y testeo a formato numérico por medio del vectorizador Tf-idf.

En el caso del modelo de clasificación para dimensiones de adaptación se adiciona un paso más por dificultades que se presentaron en el sistema de información SIIF, la dimensión 6 solo posee 2 proyectos que pueden llegar a generar un desbalance en la toma de decisiones del modelo se decide usar el proceso de Synthetic Minority Oversampling Technique o SMOTE, la cual nos permitió generar 10 proyectos sintéticos, los cuales agregamos a nuestros datos de entrenamiento y cogemos 2 aleatorios para la parte de testeo, esto se hace con el fin de reducir el desbalance de los datos para mejorar el rendimiento del modelo.

Figura 4: Ejemplo de clasificación de dimensiones por medio del modelo de aprendizaje supervisado



Fuente: Elaboración propia

Luego se implementó el modelo de aprendizaje supervisado Support Vector Machine (SVM), siendo básicamente un algoritmo que analiza los datos suministrados, llegando a visualizar patrones ocultos o estructuras intrínsecas creando fronteras de separación entre las clases en los títulos de los proyectos para su pertinente clasificación. Esta metodología se puede observar de manera más ejemplificada en la Figura 4 para el caso de las dimensiones, en el proceso de líneas estratégicas e instrumentales se sigue el mismo proceso solo que no se tiene en cuenta el paso del SMOTE ya que se presentó un buen rendimiento de clasificación por parte de los modelos sin necesidad de aplicarlo.

El rendimiento de los modelos fue aumentando o disminuyendo según la variación de los parámetros del vectorizador, como se mencionó anteriormente siendo estos el máximo de caracteres para tener en cuenta y la cantidad de n-gramas. Para el caso del modelo de aprendizaje supervisado, se tuvieron en cuenta el coste C y gamma, los cuales se comprobaron de la misma manera que en la metodología de similitud coseno en donde se iteró y se estuvo observando la matriz de confusión y las otras métricas como precision, recall, F1-Score y accuracy.

4. Resultados

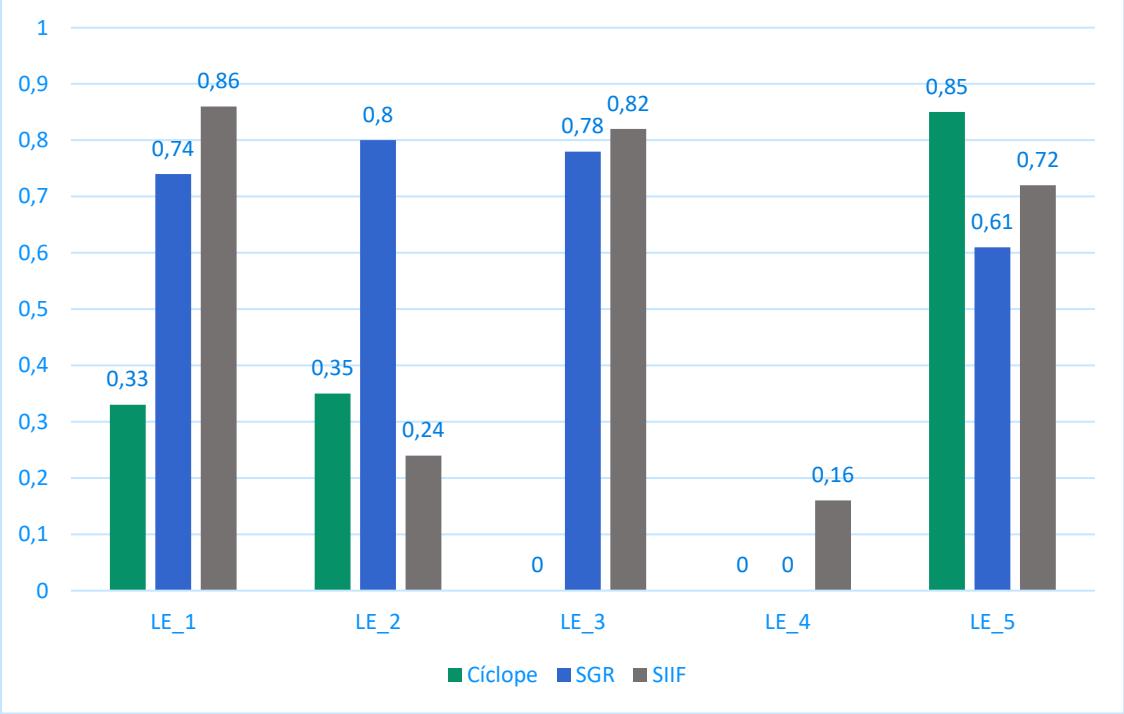
A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación. Toda retroalimentación desde un punto de vista experto o de usuario por parte de la DADS es bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

4.1. Líneas estratégicas

4.1.1. Similitud Coseno

Ahora se van a mostrar los resultados obtenidos al aplicar la similitud coseno en las líneas estratégicas para cada sistema de información, esto se hace mostrando la proporción de clasificaciones acertadas por el modelo en un rango de cero a uno. Se puede ver cuales descripciones de acciones funcionaron mejor según la línea y sistema de información.

Figura 5: Proporción de clasificación acertada en líneas estratégicas por similitud coseno

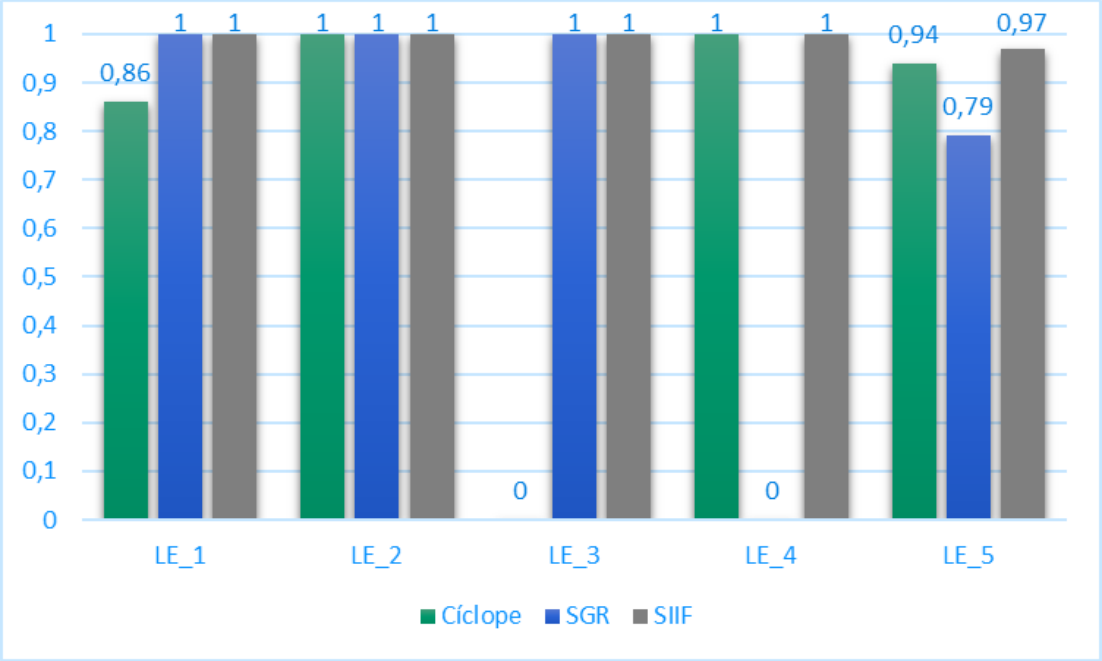


Fuente: Elaboración propia

4.1.2. Modelo de aprendizaje supervisado

Ahora se van a mostrar los resultados obtenidos al aplicar el modelo de aprendizaje supervisado en las líneas estratégicas para cada sistema de información, esto se hace mostrando la proporción de clasificaciones acertadas por el modelo en un rango de cero a uno. Se puede ver que en algunas líneas no se encuentran proyectos clasificados correctamente, esto se debe a que Cíclope no tiene proyectos pertenecientes a la línea 3 para clasificar y en la línea 4 para SGR solo se tiene un proyecto que se usó para entrenar el modelo.

Figura 6: Proporción de clasificación acertada en líneas estratégicas por modelo de aprendizaje supervisado



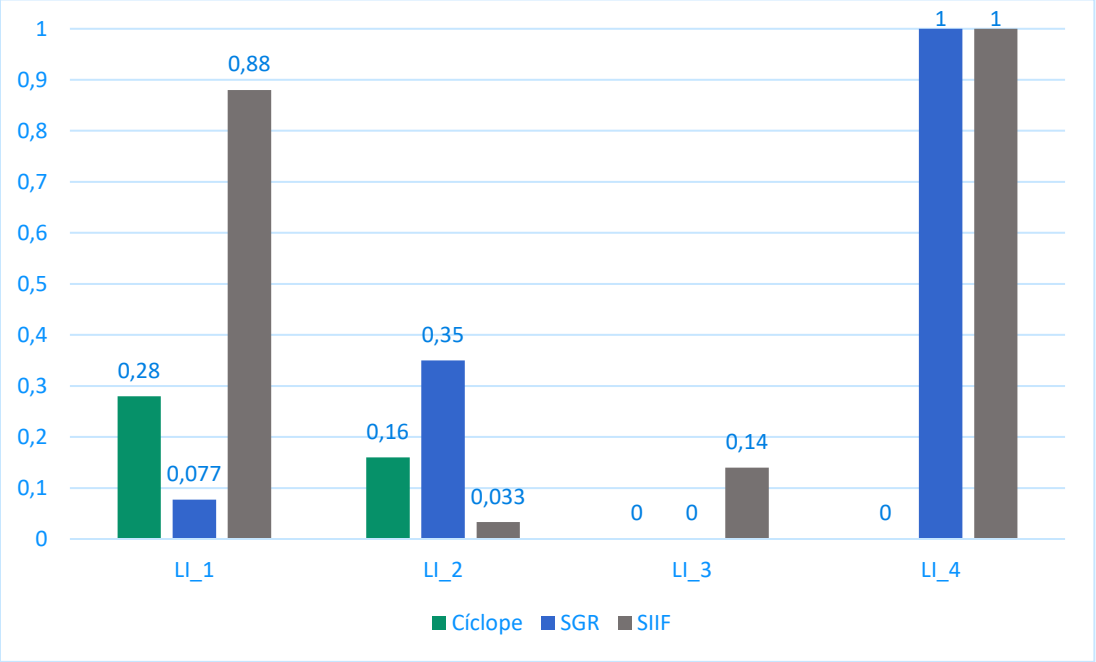
Fuente: Elaboración propia

4.2. Líneas instrumentales

4.2.1. Similitud Coseno

Se presentan a continuación de la misma manera que con las líneas estratégicas las proporciones de clasificaciones acertadas por el modelo de similitud coseno en un rango de cero a uno, aunque en el caso de las líneas instrumentales el desempeño es mucho más bajo por la poca información sobre estas líneas, en este caso no se tuvieron las descripciones de acciones de cada línea, sino que solo fueron un lineamiento para cada línea con algunos subtítulos.

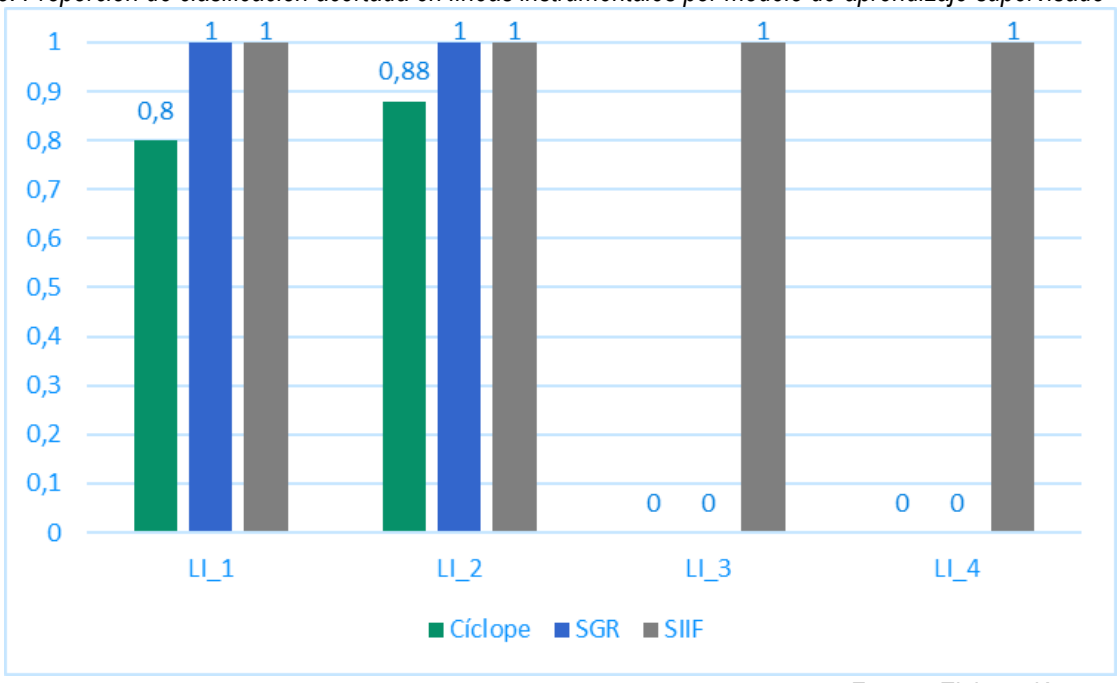
Figura 7: Proporción de clasificación acertada en líneas instrumentales por similitud coseno



Fuente: Elaboración propia

4.2.2. Modelo de aprendizaje supervisado

Figura 8: Proporción de clasificación acertada en líneas instrumentales por modelo de aprendizaje supervisado

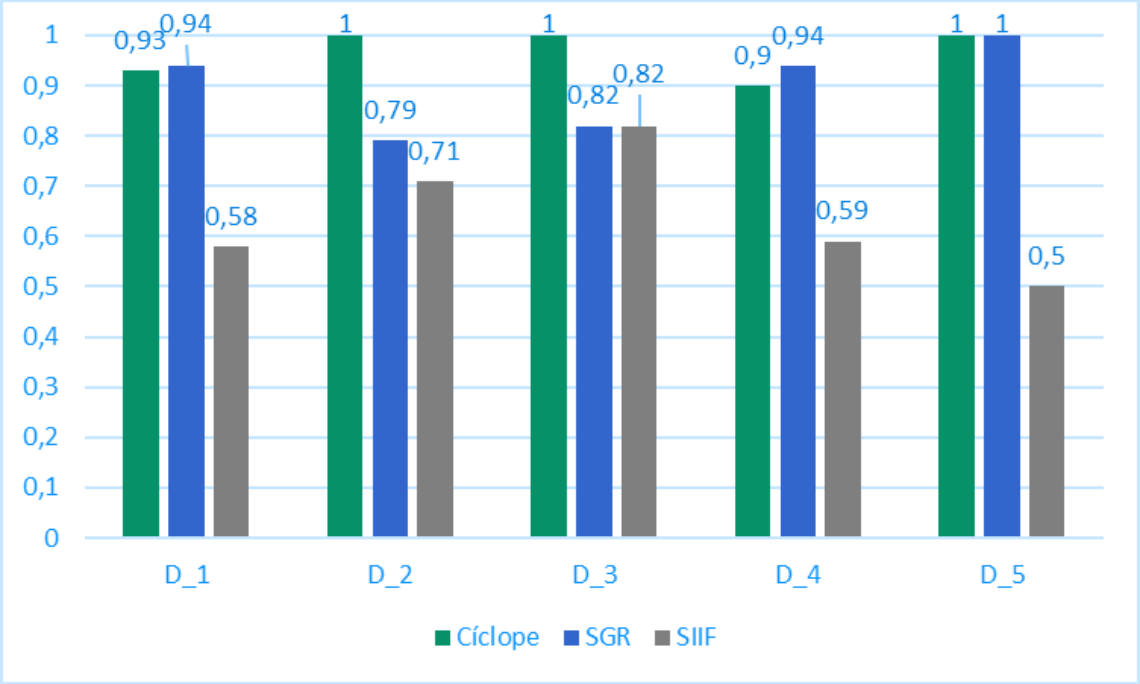


Fuente: Elaboración propia

4.3. Dimensiones

Para dimensiones se utilizó directamente la metodología del modelo de aprendizaje supervisado ya que se presentó un mejor rendimiento en comparación de la similitud coseno, en este caso se presentaron dificultades para la clasificación adecuada en el sistema de información SIIF, tratando de mejorar el desempeño en este sistema de información se crearon datos sintéticos en la dimensión cinco que era la que poseía menos datos y que por eso mismo podría conllevar un desbalance al momento de clasificar.

Figura 9: Proporción de clasificación acertada dimensiones de adaptación por modelo de aprendizaje supervisado



Fuente: Elaboración propia

5. 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. El modelo de aprendizaje supervisado presentó un mejor rendimiento para la clasificación de líneas estratégicas e instrumentales.
2. En el caso de las dimensiones de adaptación el modelo de aprendizaje supervisado presentó un buen rendimiento para los sistemas de información SGR y Cíclope.
3. Es recomendable que se aumenten los datos confiables de proyectos clasificados si se piensa en mejorar los modelos de aprendizaje supervisado o aumentar las descripciones de líneas o dimensiones de una manera más detallada si se piensa usar el modelo de similitud coseno.
4. Para SIIF es recomendable observar y analizar el motivo del que se presenten dificultades al clasificar los proyectos de este sistema de información.

6. Socialización

Los resultados del presente proyecto se socializaron con la Dirección de Ambiente y Desarrollo Sostenible DADS y la Dirección de Desarrollo Digital DDD.

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.