COLOMBIA POTENCIA DE LA VIDA

Departamento Nacional de Planeación - DNP

Dirección de Desarrollo Digital

Unidad de Científicos de Datos

Mejora regulatoria fase 4

PROYECTO DE MEJORA REGULATORIA FASE 4

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Desarrollo Digital - Unidad de Científicos de Datos • Subdirección General de Prospectiva y Desarrollo Nacional
Sector	Planeación
Tecnologías utilizadas	Python, Jinja, Google Colab
Fuentes de datos	Diario Oficial de la Imprenta Nacional de Colombia

Contenido

1. Presentación

La Política Regulatoria es un instrumento de política económica que tiene un gran impacto sobre el desarrollo de los sectores productivos y el crecimiento económico general en Colombia, principalmente la ejecución y construcción de estas políticas convergen hacia la creación de los textos regulatorios, que son la consolidación de las pautas a seguir para el correcto desarrollo e implementación de estas políticas. Por lo que es relevante comprender el nivel de complejidad tanto de la estructura de los textos, como terminología que compone su lenguaje, para entender si el nivel de complejidad de estos puede generar distorsiones en la comprensión y por lo tanto en la ejecución precisa de estas políticas, por ende, los indicadores de calidad y complejidad de textos permiten dar una aproximación a la resolución de esta pregunta. Adicionalmente, de acuerdo con las necesidades del equipo de Mejora Regulatoria, es de gran importancia, poder disponer de una información más detallada en la clasificación por sectores económicos según la CIIU (Clasificación Industrial Internacional Uniforme), por medio de una desagregación o subclasificación de cada uno de estos sectores.

De acuerdo con lo anterior, el objetivo principal de esta fase del proyecto consiste en mejorar el desempeño en el proceso de clasificación de textos regulatorios publicados en el Diario Oficial de la Imprenta Nacional de Colombia, a través de la actualización y mejora de dos modelos para tal propósito: uno para clasificar documentos regulatorios en sustanciales y no sustanciales, y el otro para clasificar documentos sustanciales por sectores. Además, de implementar una metodología semi supervisada para lograr una subclasificación de los sectores económicos y una nueva métrica de complejidad adaptada al español para medir la complejidad de los textos regulatorios. Por lo que a partir de la consecución de este proyecto se puedan implementar tales resultados como insumos que favorezcan los procesos de construcción y ejecución de la Política de Mejora Regulatoria en Colombia.

The Regulatory Policy is an instrument of economic policy that has a great impact on the development of the productive sectors and the general economic growth in Colombia, mainly the execution and construction of these policies converge towards the creation of the regulatory texts, which are the consolidation of the guidelines to follow for the correct development and implementation of these policies. Therefore, it is relevant to understand the level of complexity of the structure of the texts, as well as the terminology that composes their language, in order to understand if the level of complexity of these texts can generate distortions in the understanding and therefore in the precise execution of these policies, therefore, the indicators of quality and complexity of texts allow an approximation to the resolution of this question. Additionally, according to the needs of the "Equipo de Mejora Regulatoria", it is of great importance to have more detailed information on the classification by economic sectors according to the CIIU (International Standard Industrial Classification), by means of a disaggregation or subclassification of each of these sectors.

Accordingly, the main objective of this phase of the project is to improve performance in the process of classifying regulatory texts published in the "Diario Oficial de la Imprenta Nacional de Colombia", by updating and improving two models for this purpose: one to classify regulatory documents into substantial and non-substantial, and the other to classify substantial documents by sector. In addition, a semi-supervised methodology was implemented to achieve a sub-classification of economic sectors and a new complexity metric adapted to Spanish to measure the complexity of regulatory texts. Therefore, after the completion of this project, such results may be implemented as inputs that favor the processes of construction and execution of the Regulatory Improvement Policy in Colombia.

2. Objetivos del proyecto

2.1. General

Mejorar el desempeño de dos modelos de clasificación de documentos regulatorios, que se han implementado en fases anteriores al proyecto: uno para clasificar documentos regulatorios en sustanciales y no sustanciales, y el otro

para clasificar documentos sustanciales por sectores económicos de acuerdo con la clasificación CIIU (Clasificación industrial internacional uniforme de todas las actividades económicas).

2.2. Específicos

1. Actualizar el clasificador existente para los documentos normativos según la categoría sustancial, con la nueva información etiquetada compartida por el Grupo de Mejora Regulatoria.
2. Actualizar el clasificador existente de los documentos normativos según los sectores productivos CIIU, con la nueva información etiquetada compartida por el Grupo de Mejora Regulatoria.
3. Implementar una metodología no supervisada que logre obtener mayor desagregación para ciertos sectores económicos, que pueden estar compuestos por varios subsectores.
4. Implementar una nueva métrica de complejidad para los textos normativos, de tal manera que corresponda con una metodología más cercana al lenguaje natural de la lengua española, con el cual está redactada la normativa nacional.
5. Actualización de la clasificación para nuevos documentos normativos con el corte más actual disponibles en el servidor del diario oficial.

3. Metodología

Es importante destacar que el clasificador utilizado en esta fase es el mismo que se usó en fases anteriores, dado que el proceso de selección de éste resultó de un proceso riguroso, que consistió en probar varios clasificadores, y se determinó que Support Vector Machine (SVM) fue el que ofreció el mejor desempeño, tanto para la clasificación sustancial como para la clasificación por sectores económicos. Por esta razón, se decidió utilizar SVM para los procesos de reentrenamiento y reajuste de parámetros, con el objetivo de mejorar aún más su desempeño y lograr resultados aún más precisos y confiables.

En la primera etapa de esta fase, se obtuvo un primer modelo para la clasificación por sectores de los documentos clasificados como “Sustancial” mediante el entrenamiento con la información total etiquetada manualmente, el cual mejoró en la precisión hasta 4 puntos porcentuales en promedio, comparado con el promedio de precisión obtenido en fases anteriores. Dado que, en las etapas posteriores, específicamente para el proceso de subclasificación, se dio la necesidad de implementar una nueva metodología que permita mejorar aún más este desempeño, esta última es la que se consolidó, por lo que este informe se centra en explicar en detalle la metodología y los resultados obtenidos del modelo que presentó mejor desempeño.

En esta sección se describe en detalle la metodología empleada en el proceso de ajuste y mejoras del modelo de clasificación automática de los 10 sectores económicos[1] para lograr una clasificación más precisa de los documentos regulatorios clasificados como sustanciales. Asimismo, se presentan los detalles de la implementación de la metodología semi supervisada utilizada para lograr un mayor grado de desagregación de estos 10 sectores en subsectores más específicos, de acuerdo con la subclasificación mostrada en la Tabla 6, del Anexo 1. Por último, se abordan los detalles de implementación de la métrica SSR (Spaulding’s Spanish Readability), el visualizador interactivo de los resultados y la herramienta de tipo Notebook que resume todo el proceso de clasificación para los documentos.

3.1. Actualización de los modelos de clasificación

Con el objetivo de mejorar significativamente el desempeño en la clasificación automática de documentos normativos sustanciales dentro de los 10 sectores, se implementó una metodología que mejoró el modelo de clasificación,

[1] Sectores económicos CIIU (Clasificación Industrial Internacional Uniforme): 1. Agricultura, caza, forestal y pesca, 2. Minería y extracción, 3. Manufacturas, 4. Electricidad, gas y suministros de agua, 5. Construcción, 6. Comercio, hoteles y restaurantes, 7. Transporte, almacenaje y comunicaciones, 8. Finanzas, negocios y bienes raíces, 9. Servicios personales, administración pública, salud, educación, 10. Otros (Administrativos)

considerando que las predicciones de este son el punto de partida para la etapa de desagregación en subsectores más específicos (ver Tabla 7, del Anexo 1).

La metodología de mejora se centró en aplicar una reducción de dimensionalidad a los vectores de frecuencia de palabras obtenido de los textos de entrenamiento, utilizando la técnica T-SVD (Truncated - Singular Value Decomposition), la cual permite eliminar características redundantes y no esenciales de los vectores de palabras que podrían afectar el desempeño del clasificador. Específicamente, en matrices generadas por vectorizadores de texto, podría haber variables fuertemente correlacionadas que contienen ruido y complejidad, lo que podía dificultar la obtención de un buen clasificador. Por lo tanto, la reducción de dimensionalidad mediante T-SVD ayuda a eliminar estas redundancias y mejorar la calidad de los datos de entrada al clasificador SVM. Este enfoque permitió un mejor aprovechamiento de las características más relevantes para la clasificación y, en consecuencia, una mejora significativa en la precisión y eficiencia del modelo, cuyo promedio pasó del 64% al 68%. En la Figura 1, se presenta el flujo general de la nueva metodología de clasificación de sectores.

Figura 1: Procedimiento implementado para obtener el modelo de clasificación.

Fuente: Elaboración propia

3.2. Metodología semi supervisada para la desagregación de los 10 sectores económicos

En el proceso de desagregación, a diferencia de la metodología utilizada en la clasificación automática de sectores del apartado anterior, no se dispone de datos previamente etiquetados para obtener esta subclasificación de cada sector (ver *Tabla 7* del Anexo 1), además, intentar implementar un nuevo proceso de etiquetado manual con ayuda del Grupo de Mejora Regulatoria (GMR), para los 43 diferentes subsectores, resultaría extremadamente tedioso y requeriría una cantidad significativa de tiempo para obtener suficientes datos de entrenamiento.

Para abordar el desafío de obtener datos de entrenamiento para cada subsector de manera *semi-supervisada*, se adoptó una estrategia basada en búsqueda de términos clave. En colaboración con el GMR, se ha desarrollado un diccionario de palabras clave específico para cada subsector, de manera que cada uno de ellos cuenta con un conjunto de términos que lo identifican de manera única (Anexo 2). En la *Tabla 1* se ejemplifica el diccionario en mención para un caso particular (Sector 2, Subsector 1).

Tabla 1: Ejemplo del diccionario de palabras para un caso particular.

Sector	Subsector	Palabras primarias	Palabras activadoras
2	1	extracción, extracción y aglomeración, explotación	hulla, carbón de piedra, carbón lignítico, turba, carbón

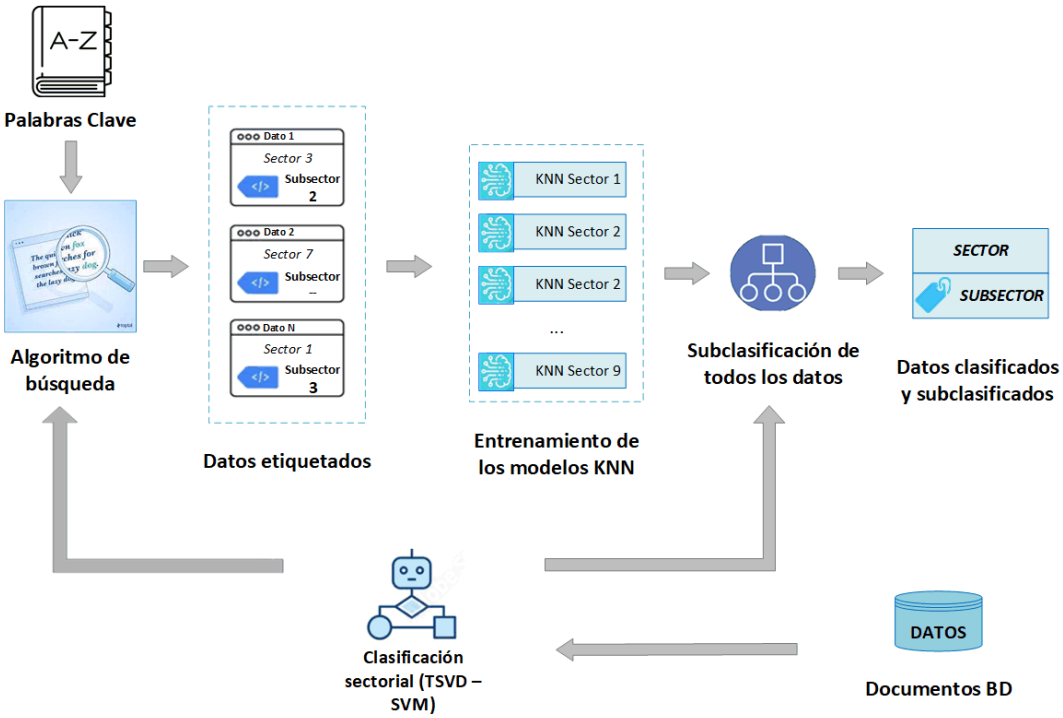
Fuente: Elaboración propia

Este diccionario se ha estructurado en dos conjuntos de palabras clave: las palabras primarias y las palabras activadoras. Las palabras primarias son aquellas que son altamente representativas y específicas de cada subsector, reflejando sus características distintivas. Por otro lado, las palabras activadoras son términos que, cuando se encuentran relativamente cercanas a una palabra primaria en el texto de un documento, indican una alta probabilidad de que dicho documento esté relacionado con un subsector específico. Por ejemplo, para el caso particular presentado en la *Tabla 1*, si en un determinado documento previamente clasificado en el sector 2, el algoritmo encuentra la palabra "extracción", y muy cerca a esta se encuentra la palabra "carbón lignítico", éste será etiquetado en el Subsector 1 del Sector 2.

El proceso de clasificación se realiza de la siguiente manera: si en al menos una sección del texto de un documento se encuentran tanto una palabra primaria como una palabra activadora de un determinado subsector, el algoritmo clasifica automáticamente el documento como perteneciente a ese subsector. En el caso en que el algoritmo no logre encontrar ninguna combinación de palabras clave correspondiente a algún subsector, la etiqueta de subsector del documento quedará vacía.

Al aplicar el algoritmo de palabras clave por sector, se logra etiquetar un conjunto de documentos que se convierten en los datos de entrenamiento. Estos documentos etiquetados se utilizan como insumo para entrenar los modelos de clasificación de cada subsector mediante el algoritmo *K-Nearest Neighbors (KNN)*. El procedimiento completo se muestra de manera esquemática en la *Figura 2*.

Figura 2: Procedimiento implementado para la subclasificación de los sectores económicos.



Fuente: Elaboración propia

En el proceso de desagregación, se crean 8 modelos KNN, uno para cada sector económico predefinido. Es importante destacar que para el sector 5 y 10, no se requiere la construcción de un modelo de desagregación, ya que no cuentan con subsectores. La singularidad de estos modelos radica en que el número k de vecinos más cercanos se adapta según la cantidad de subsectores que pueda tener cada sector. Por ejemplo, en el caso del Sector 1, para su respectivo modelo KNN, el número de vecinos más cercanos se ajusta a 3, ya que este es el número de subsectores asociados a dicho sector. Cada modelo KNN se entrena con datos de entrenamiento específicos correspondientes a su respectivo sector, los cuales fueron etiquetados utilizando el algoritmo de palabras clave por sector.

Estos modelos KNN de cada sector, aprenden de los datos de entrenamiento provenientes del algoritmo de búsqueda, y utilizan la información de los subsectores presentes en dichos datos para realizar clasificaciones precisas para todos los documentos presentes en la base de datos. El proceso de clasificación, de cada modelo KNN para un dato específico (documento), consiste inicialmente en calcular las distancias existentes entre el dato nuevo y los datos de entrenamiento, luego, con estas distancias, selecciona los k vecinos más cercanos, y dentro de estos selecciona los datos cuya categoría (Subgrupo) sea la más relevante, la cual será el subgrupo del nuevo dato.

3.3. Métrica de complejidad de lectura SSR

La métrica de complejidad *Spaulding's Spanish Readability (SSR)* es una herramienta que permite evaluar el nivel de dificultad de un texto en español. Esta métrica se basa en el análisis de tres factores: la cantidad promedio de palabras por oración en el texto analizado; la cantidad promedio de palabras desconocidas y un término de ajuste.

$$SSR = 1.609 * (Palabras\ por\ Oración) + 331.8 * (Promedio\ palabras\ desconocidas) + ajuste$$

$$SSR = 1.609 * \frac{|w|}{|s|} + 331.8 * \frac{|rw|}{|w|} + \beta$$

Donde: $|w|$ representa el número de palabras en el texto, $|s|$ el número de oraciones; $|rw|$ el número de palabras desconocidas, que son aquellas que no aparecen en el listado de palabras frecuentes del idioma español del corpus CREA de la Real Academia Española (RAE) y β que es una constante de ajuste que por defecto es 22.

El puntaje obtenido de implementar la métrica SSR se puede analizar de manera categórica tal como se muestra en la *Tabla 2*.

Tabla 2: Interpretación del puntaje que se obtiene de aplicar la métrica SSR.

Puntaje SSR	Facilidad de lectura
Menor a 40	Material muy simplificado
40 - 60	Muy fácil
61 - 80	Fácil
81 - 100	Dificultad moderada
101 - 120	Difícil
121 o más	Muy difícil

Fuente: Elaboración propia

El termino de ajuste β puede ser aumentado o disminuido dado un conocimiento a priori de los textos evaluados o del contexto en el cual se analiza la complejidad de los textos.

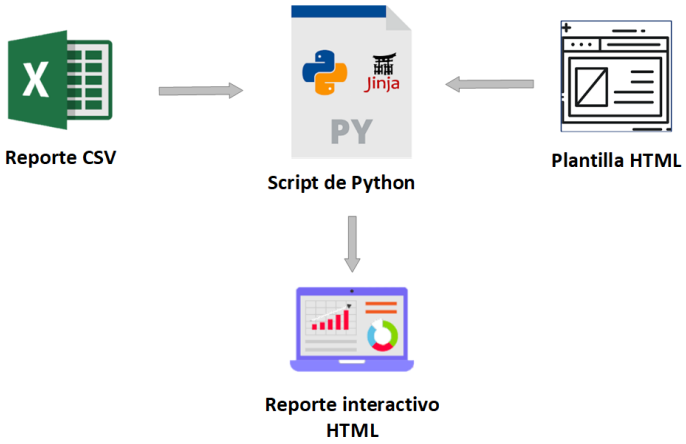
3.4. Actualización del reporte interactivo para visualización de resultados

El objetivo del reporte interactivo es facilitar la consulta y el entendimiento de los resultados para facilitar la obtención de conclusiones referente tanto a las métricas del grado de complejidad en la lectura de textos como a la clasificación en los sectores y subsectores económicos, donde se concentra la literatura regulatoria, con el principal propósito de poder brindar insumos que apoyen los procesos de formulación y vigilancia de la política de mejora regulatoria en Colombia.

Partiendo que de fases anteriores ya se ha realizado este aplicativo para visualización de resultados, para esta fase lo que se busca es poder actualizarlo, dado que en el transcurso del tiempo se han generado cambios, lo que obliga a cambiar la estructura del reporte. Como primera actividad en la metodología, se realizan los ajustes necesarios a la plantilla del reporte.

En la Figura 3 se muestra de manera resumida, el procedimiento implementado para generar el correspondiente reporte interactivo de tipo HTML. Como se puede observar, todo el proceso es realizado por medio de un script de Python, el cual contiene el código necesario para este propósito. Como información de entrada, se tiene el reporte de tipo CSV, el cual contiene todos los datos de salida como resultado del procedimiento de clasificación implementado en esta fase, y los cálculos de métricas de complejidad tanto de esta fase como de fases anteriores. Como insumo de entrada, también se tiene la plantilla del reporte previamente actualizada y ajustada para contener la nueva información generada en esta fase.

Figura 3: Metodología para la obtención del reporte interactivo



Fuente: Elaboración propia

El script, inicialmente toma la información contenida en el reporte CSV y la organiza de acuerdo con la estructura requerida en la plantilla, calculando información adicional necesaria, así como los arreglos de información para cada una de las diferentes gráficas que se presentarán en el reporte interactivo. Luego, por medio de una librería denominada Jinja2, manipula la plantilla HTM y le entrega cada una de las secciones de información previamente organizada, para presentarla y graficarla de acuerdo a la forma y a los estilos allí establecidos. Enseguida, con la información contenida en la plantilla, el script procede a generar un nuevo archivo con el nombre "Reporte.html", el cual corresponde la herramienta interactiva para la visualización de los resultados.

3.5. Herramienta de tipo Notebook

El principal propósito de esta herramienta es permitir ejecutar de manera sencilla, todo el procesamiento, clasificación y cálculo de métricas, de los documentos normativos desarrollado en las diferentes fases del proyecto, con el principal objetivo de poder actualizar el reporte general, y el reporte de consulta de resultados.

La herramienta Notebook se desarrolló con Google Colab, el cual es un producto de Google Research, que permite escribir y ejecutar código de Python desde el navegador, sin la necesidad de instalar software adicional. Esta herramienta, contiene los pasos necesarios para el procesamiento de los documentos regulatorios de interés, con sus respectivas instrucciones, por lo tanto, su propósito principal es la actualización del reporte general que se ha venido consolidando con el desarrollo de este proyecto, con los nuevos documentos regulatorios que se van a ir generando en el futuro.

El proceso que ejecuta la herramienta básicamente consiste en la carga y procesamiento de los documentos normativos, en la clasificación de estos documentos, tanto sustancial como por Sectores Económicos, la clasificación de estos sectores en subsectores, así como el cálculo de las diferentes métricas que permiten estimar el grado de dificultad en la lectura de estos documentos. Adicionalmente, mediante esta herramienta se busca poder actualizar el respectivo reporte interactivo, de tal manera que se pueda incluir la información futura en la visualización.

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 del GMR es bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

A continuación, se presentan los resultados obtenidos, los cuales son presentados en 5 subsecciones, para tener mayor detalle y mayor claridad, debido a la extensión del contenido de este documento. Cada subsección presenta una breve explicación inicial, del procedimiento que se implementó para lograr estos resultados.

4.1. Descripción de la base de datos

Inicialmente, es importante tener en cuenta que los documentos normativos que conformaron la base de datos utilizada como fuente de información principal para este entregable, corresponden a todos los documentos existentes en el diario oficial, a partir del año 1991 hasta junio de 2023, para un total de 73.903 documentos procesados, de los cuales 52.416 documentos fueron clasificados en la categoría *Sustancial*.

4.2. Modelo de clasificación sustancial

Para entrenar el modelo de clasificación sustancial – no sustancial, se contó con una base de datos total de 2.043 documentos etiquetados manualmente por el Grupo de Mejora Regulatoria del DNP, de los cuales 452 corresponden a la nueva validación entregada. De este total, 1.236 documentos se etiquetaron como sustanciales (clase 1) y 807 documentos etiquetados como no sustanciales (clase 0).

En la Tabla 3 se presentan los resultados del desempeño individual del clasificador por cada clase en 307 muestras de prueba, los cuales muestran que se obtuvo una precisión[2] promedio estimada de 89%, un *recall*[3] promedio de 86%, que, al analizarlo por clase, se evidencia un mayor ajuste del clasificador en la clase (1) (96%), es decir, una mayor confiabilidad al clasificar un documento como *Sustancial*, en comparación al 76% de la clase (0). Por lo tanto, el clasificador permite cometer mayor tasa de Falsos positivos (clasificar como sustancial un documento que no lo es), lo que minimiza el riesgo de dejar por fuera una normativa sustancial. Con estos resultados, podemos inferir que el desempeño del modelo actualizado mejoró 7 puntos porcentuales en la precisión promedio, comparado con la precisión obtenida en fases anteriores y 5 puntos porcentuales en el *recall* promedio.

Tabla 3: Resultados de desempeño del clasificador SVM para cada una de las clases: (0) no sustancial, (1) sustancial

Clase	Precisión	Recall	F1 - score	N. muestras prueba
0 – No sustanciales	92%	76%	83%	129
1 - Sustanciales	85%	96%	90%	178
Promedio	**89%**	**86%**	**87%**	**307**

Fuente: Elaboración propia

4.3. Modelo de clasificación de sectores económicos

Para el entrenamiento del modelo de sectores, se empleó la base de datos que consta de 1.035 registros previamente validados y etiquetados manualmente por el GMR para la clasificación automática de los 10 sectores. El proceso de entrenamiento, ilustrado en la Figura 1, implica iteraciones múltiples con los datos de entrenamiento para obtener los parámetros óptimos tanto para el transformador TSVD (número de componentes o dimensiones) como para el modelo de clasificación SVM.

Luego de llevar a cabo el proceso de entrenamiento, se determinaron los mejores parámetros para ambos componentes (ver Tabla 4). Estos parámetros resultaron ser cruciales para el rendimiento óptimo del modelo. El transformador T-SVD y el modelo de clasificación SVM demostraron un rendimiento sobresaliente al ajustar los parámetros de acuerdo con los datos de entrenamiento.

[2] Capacidad del modelo de dar el resultado deseado con exactitud.
[3] También conocido como el ratio de verdaderos positivos, es utilizado para saber cuántos valores positivos son correctamente clasificados.

Tabla 4: Parámetros con los cuales se obtiene el mejor desempeño del modelo de clasificación.

TSVD	SVM
Algorithm = randomized **n_components** = 300	**C**= 2.2 **gamma**= 0.8 **Kernel** = "rbf" **class_weight** = "balanced"

Fuente: Elaboración propia

Con el propósito de poder disponer de una comparación detallada, se presentan los resultados de desempeño tanto del modelo obtenido inicialmente en esta fase (Tabla 5), como del nuevo modelo (Tabla 6), los resultados por sector en los que el modelo supera al anterior son presentados en "Negrita", y los resultados inferiores son resaltados. Los resultados de desempeño obtenidos indican que este nuevo modelo obtuvo en promedio un 68% en *Precisión*, un 65% en *Recall*, lo cual, en comparación con el promedio de precisión obtenido en el modelo inicial de esta fase del proyecto, refleja una mejora significativa de 4 puntos porcentuales en precisión promedio, y de 5 puntos porcentuales en *Recall* promedio.

Tabla 5: Resultados de desempeño del clasificador SVM anterior, para la clasificación sectorial

Clase	Precisión (%)	Recall (%)	F1 – score (%)	N. muestras prueba
1	58	64	61	11
2	100	80	89	10
3	73	89	80	18
4	75	60	67	5
5	67	44	53	9
6	36	29	32	14
7	57	62	59	13
8	62	67	64	24
9	45	50	47	26
10	62	58	60	26
Promedio	**64%**	**60%**	**61%**	**156**

Fuente: Elaboración propia.

Tabla 6: Resultados de desempeño del clasificador SVM actual, para la clasificación sectorial.

Clase	Precisión (%)	Recall (%)	F1 – score (%)	N. muestras prueba
1	**86**	55	**67**	11
2	90	**90**	**90**	10
3	**81**	**94**	**87**	18
4	*75*	*60*	*67*	5
5	40	*44*	42	9
6	**60**	**43**	**50**	14
7	56	**69**	**62**	13
8	**65**	**71**	**68**	24
9	**56**	**58**	**57**	26
10	**68**	**65**	**67**	26
Promedio	**68%**	**65%**	**66%**	**156**

Fuente: Elaboración propia.

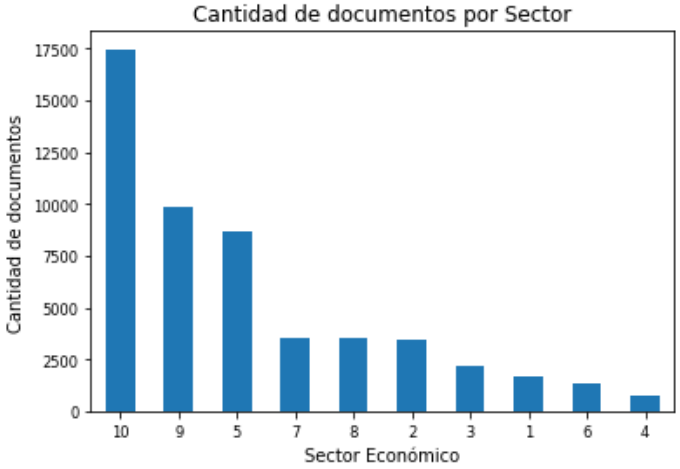
Comparando los resultados por cada sector, con el modelo anterior, en cuanto a precisión, el nuevo modelo presenta mejores o iguales resultados en la mayoría de las categorías, excepto en los sectores 2, 5 y 7, en los cuales el sector

obtiene porcentajes inferiores. Es importante destacar que las clases 5, 7 y 9 presentaron los porcentajes más bajos de precisión para este nuevo modelo, con 40%, 56% y 56% respectivamente, mientras que las clases 2, 3 y 8 exhibieron los mejores resultados en términos de *Recall*, con un 90%, 94% y 71% respectivamente.

Estos resultados reflejan el éxito del proceso de entrenamiento y optimización del modelo SVM. La mejora en la precisión del modelo proporciona una mayor confianza en su capacidad para clasificar adecuadamente los documentos en sus respectivos sectores económicos. El hecho de que algunas clases obtengan un alto porcentaje de *Recall* también indica la eficacia del modelo en identificar de manera precisa y exhaustiva los documentos relacionados con esos sectores específicos.

Naturalmente, luego de obtener el modelo ajustado a sus parámetros óptimos, se procede a la fase de predicciones en la clasificación sectorial, esta clasificación en los diferentes sectores económicos servirá como punto de partida para la siguiente etapa, correspondiente a la desagregación de la clasificación sectorial. En la Figura 4 se puede observar una distribución de la cantidad de documentos que el modelo clasificó por sector. De los 52.416 documentos normativos clasificados como *Sustanciales*, el sector con mayor cantidad de documentos clasificados, con un total de 17.474 (33.3%), es el número 10 correspondiente al sector *Otros (Administrativo)*, y el que menos cantidad presenta, con un total de 727 (1.4%), es el número 4, correspondiente al sector de *Electricidad, gas y suministros de agua*.

Figura 4: Distribución de las predicciones hechas por el modelo automático por sector económico.



Fuente: Elaboración propia.

4.4. Desagregación de la clasificación sectorial

La Figura 5 presenta los resultados tras la implementación del algoritmo de búsqueda, el cual se basa en el diccionario de palabras primarias y activadoras detallado en la Subsección 3.2. Esta gráfica ilustra la frecuencia de asignación a los subsectores[4], considerando su clasificación previa en los 10 sectores económicos. Es importante destacar que el sector 5 y 10 no cuentan con subsectores, por lo tanto, en la figura no se muestra información correspondiente a estos.

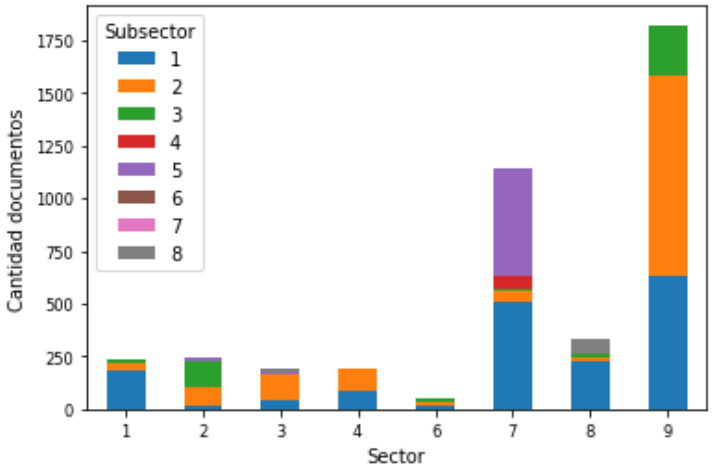
De los 52.416 documentos clasificados como Sustanciales y distribuidos en sus respectivos sectores económicos mediante el modelo presentado en la Sección 4.3, únicamente 4.207 documentos pudieron ser etiquetados mediante el algoritmo de búsqueda de palabras clave, lo que representa aproximadamente un 8%. Es importante destacar que la distribución desbalanceada de la asignación de sectores, ilustrada en la Figura 4, ejerce una influencia directa en el

[4] En la *Figura 5*, los subsectores se encuentran diferenciados por colores, pero es importante tener en cuenta que la categorización de cada subsector varía según su respectivo sector. En otras palabras, la clasificación de un documento en el Subsector 1 (Agricultura, ganadería, caza y actividades de servicios conexas) del Sector 1 difiere de la clasificación de un documento en el Subsector 1 (Extracción de carbón, carbón lignítico y turba) del Sector 2.

número de documentos que serán clasificados en cada subsector dentro de un sector económico. Este fenómeno se refleja en el hecho de que el sector 9 cuenta con una mayor cantidad de muestras para el entrenamiento de subclasificación, sumando un total de 1.821 documentos subclasificados. En contraste, el sector 6 presenta una frecuencia significativamente menor, con tan solo 54 documentos subclasificados utilizados para el entrenamiento de los modelos KNN.

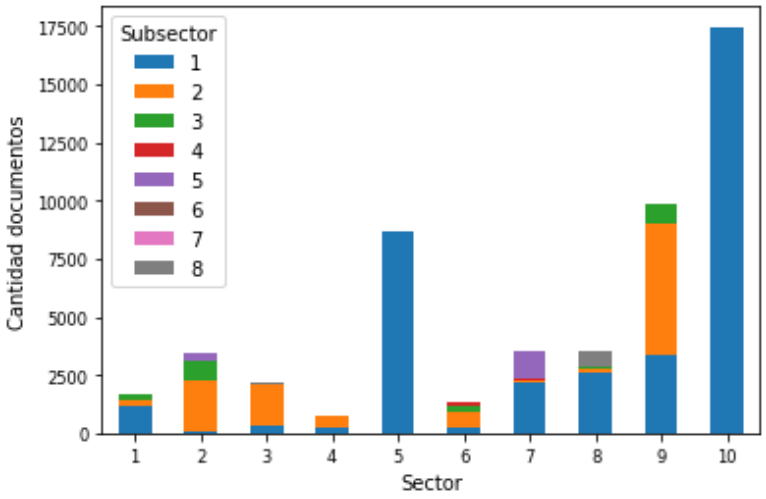
Tal como se expuso en la Subsección 3.2, para cada conjunto de documentos clasificados en los 8 sectores (excepto los sectores 5 y 10), se generó un modelo de subclasificación KNN propio. Cada uno de estos modelos fue entrenado utilizando los datos etiquetados por el algoritmo de búsqueda correspondientes a su sector respectivo, mostrados en la Figura 5.

Figura 5: Distribución de documentos etiquetados por sector y subsector.



Fuente: Elaboración propia.

Figura 6: Distribución de documentos clasificados por los modelos KNN, por sector y subsector



Fuente: Elaboración propia

Después de ajustar los modelos KNN con la información obtenida a través de la estrategia de palabras clave, se procede a realizar las desagregaciones por subsector de los 52.416 documentos clasificados como Sustanciales,

considerando su clasificación dentro de los 10 sectores económicos. En la Figura 6, se ilustra la distribución de los documentos subclasificados por medio de los modelos KNN. Se aclara nuevamente que los sectores 5 y 10 no cuentan con subclasificación. Como se esperaba, la distribución de los resultados de la subclasificación, tal como se muestra en la Figura 6, refleja un comportamiento muy similar a los datos de entrenamiento previos (Figura *5*).

Además de la etiqueta de subclasificación, se obtiene también la probabilidad con la que cada documento pertenece al subsector etiquetado, esto para que el equipo de expertos pueda analizar de mejor manera la pertenencia o no de los documentos en los diferentes subsectores. Es importante aclarar, que los documentos con probabilidad de pertenencia igual a 1, son los documentos de entrenamiento que el algoritmo de búsqueda etiquetó previamente, por lo que el modelo en su predicción, lo definirá con una probabilidad 100% de pertenencia al subgrupo.

4.5. Métrica de complejidad SSR

La métrica SSR fue aplicada inicialmente a una muestra aleatoria de 400 documentos normativos seleccionados de la base de datos. Estos documentos fueron distribuidos en cuatro grupos de 100 muestras cada uno, según el año de emisión. El primer grupo incluyó documentos emitidos entre los años 1991 y 2014, el segundo grupo abarcó los años 2015 a 2019, el tercer grupo comprendió los años 2020 y 2021, y finalmente, el cuarto grupo consistió en muestras del año 2022. Esta división por rangos de años fue una decisión tomada en conjunto con el GMR.

El objetivo de esta división fue examinar si existía alguna tendencia en la complejidad de los textos a lo largo de los años, evaluando el *score* obtenido por cada grupo en la métrica SSR. Sin embargo, los resultados demostraron que no existían diferencias significativas en la complejidad de los documentos entre los distintos grupos. En promedio, cada grupo obtuvo una categoría de dificultad "*Difícil de entender*" según la clasificación de la métrica.

Después de analizar los resultados, se determinó que se requería un ajuste para todos los intervalos. Para ello, se llevó a cabo una clasificación manual de algunos documentos en la categoría en la que deberían estar a criterio de los expertos del GMR, tomando estos documentos como referencia para calcular el ajuste correspondiente. El proceso consistió en calcular las diferencias entre los puntajes obtenidos inicialmente para estas muestras y los puntajes que deberían obtener realmente. A partir de estas diferencias, se seleccionó un valor de ajuste que permitiera que la mayoría de los documentos de referencia se ajustaran a la categoría correcta.

El ajuste obtenido fue una reducción de 12 puntos para todos los intervalos. Por lo tanto, el término de ajuste en la fórmula de la métrica SSR pasó de 22 a 10. La fórmula ajustada es la siguiente:

$$SSR = 1.609 * \frac{|w|}{|s|} + 331.8 * \frac{|rw|}{|w|} + 10.0$$

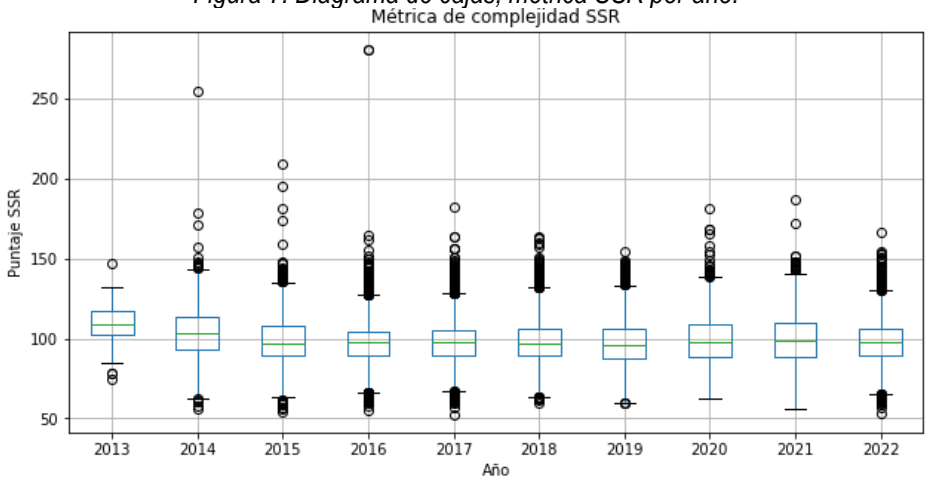
Al implementar la métrica de complejidad a todo el conjunto de datos, es decir, para los documentos normativos existentes en el diario oficial, desde 1991 hasta 2023, para un total de 73.903 documentos procesados, los resultados muestran que la métrica *SSR* obtenida, en promedio se ubica en un valor de 96.8, el cual está dentro de la categoría *Dificultad Moderada* de acuerdo con la *Tabla 2*, el 50% de los datos se ubica en un valor de 96.3 o menos, el cual está dentro de la misma categoría del promedio. El 25% de los datos, presenta una métrica de 87 puntos o menos, que al igual que el promedio y la media, está dentro de la categoría de *Dificultad moderada* o a una categoría de menor dificultad.

La Figura 7 muestra el comportamiento de las medidas de tendencia central para los documentos normativos publicados en los últimos 10 años[5]. Los resultados indican que la media del puntaje *SSR* se ubica en un valor entre 90

[5] Estos períodos seleccionados corresponden a los últimos 10 años inmediatamente anteriores al año 2023, pues al momento de la elaboración de este informe, no se tiene la información completa de todos los meses del mismo, por lo tanto, no se lo considera en este análisis por años, pero sus datos si son considerados para las medidas estadísticas de toda la información presente hasta el momento.

y 100 para todos los años, excepto los años 2013 y 2014, en los que la media supera los 100 puntos. Es decir que, en la gran mayoría de los últimos 10 años, el 50% de los documentos fueron clasificados en la categoría *Dificultad Moderada* o en una categoría de dificultad menor. Se puede observar también, que el año 2013 es el que presenta la media más alta con un valor de 109, mientras que el año 2019 presenta la media más baja, con un valor de 95.6.

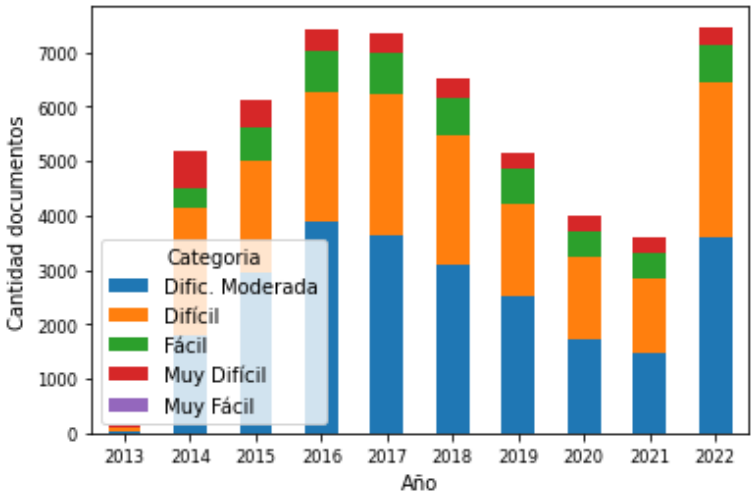
Figura 7: Diagrama de cajas, métrica SSR por año.



Fuente: Elaboración propia

De igual manera, en la Figura 8 se muestra la distribución de la cantidad de documentos calificados según esta métrica, por categoría, para cada uno de estos últimos 10 años. Los resultados muestran que, para la mayoría de estos años, excepto 2013 y 2014, la categoría más frecuente es la de *Dificultad Moderada* como se esperaba, debido al ajuste que se realizó en la métrica, la segunda categoría más frecuente fue la de *Difícil*, y la categoría menos frecuente fue la de *Muy Difícil*.

Figura 8: Distribución de la frecuencia de documentos calificados por año y por categoría

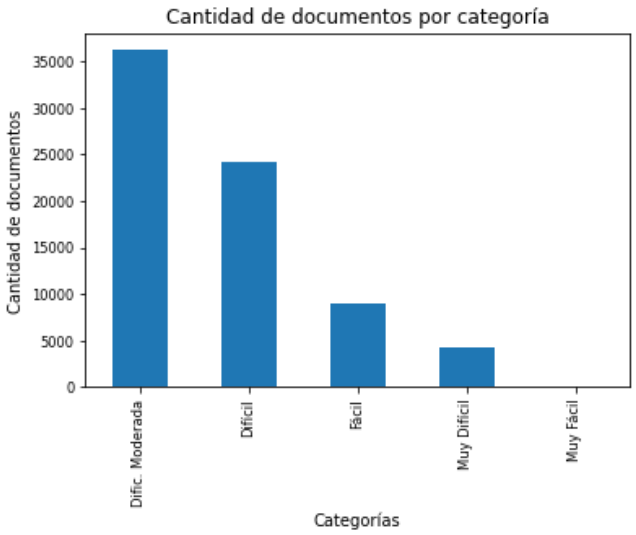


Fuente: Elaboración propia.

En la Figura 9 podemos observar la frecuencia por categoría con la que los documentos fueron clasificados según la métrica SSR. Cómo se observó con las medidas de tendencia central, la gran mayoría de los documentos fueron clasificados en la categoría *Dificultad Moderada*, lo que corresponde al 49.1% de los documentos, esto debido al ajuste realizado en la métrica. Vemos también que la categoría con la segunda mayor frecuencia es la *Difícil* con un porcentaje

correspondiente al 32.8%, y la categoría que presenta la menor frecuencia es la *Muy Fácil,* con un porcentaje del 0.1%. El 12.2% de los documentos, fueron calificados en la categoría *Fácil,* y el 5.8% de los documentos se calificaron como *Muy Difícil* en su grado de lectura según la métrica SSR.

Figura 9: Diagrama de barras, métrica SSR por categoría.



Fuente: Elaboración propia

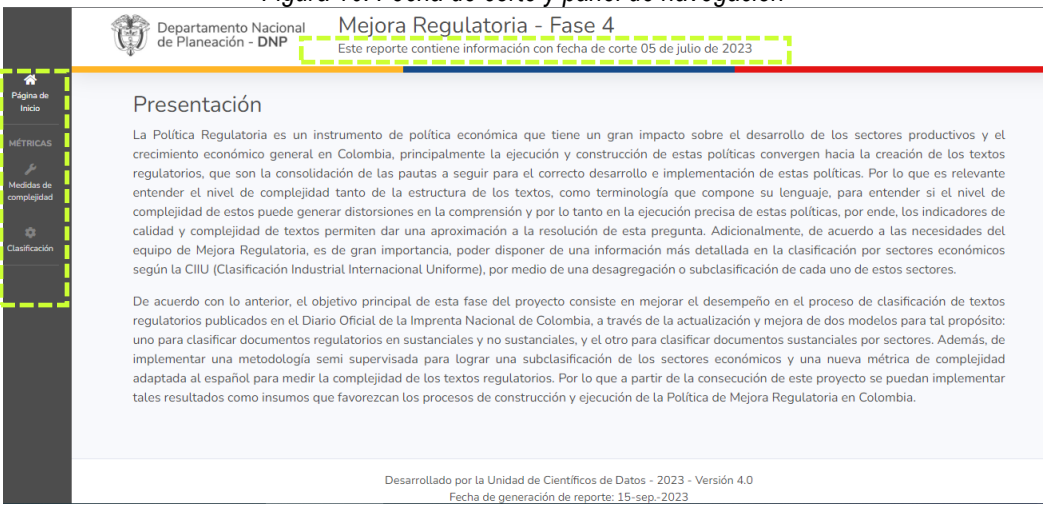
4.6. Reporte interactivo para visualización

Tal como se mencionó en la metodología, el propósito del reporte interactivo es facilitar la consulta y el entendimiento de los resultados obtenidos de todo el proceso realizado a los documentos regulatorios. Para esta fase, el reporte fue actualizado en 4 aspectos principales: el cambio de estilo en la presentación, inclusión de la fecha de corte, se agregó la sección para visualización de los resultados de la nueva métrica de complejidad SSR, y la inclusión de la sección para visualizar los resultados de la subclasificación de los diferentes sectores económicos.

El reporte desarrollado, contiene 3 secciones principales: la página de inicio, métricas de complejidad y los resultados de la clasificación. En la Figura 10 se muestra una captura de pantalla de la página de inicio del reporte interactivo desarrollado, en cuya parte superior, justo debajo del título se muestra la fecha de corte de la información presentada en dicho reporte, en la parte central de la ventana se presenta una definición sobre la política de mejora regulatoria, así como una explicación breve sobre metodología que se implementó para procesar los textos de la política regulatoria y el objetivo principal del proyecto. En la parte lateral izquierda se ubica el panel de navegación del reporte, el cual contiene varios botones que permiten la selección de las diferentes secciones que contienen los resultados.

La sección “Medidas de complejidad”, contiene los resultados de las diferentes métricas implementadas en fases anteriores incluida la métrica SSR implementada en esta fase, para cada uno de los años disponibles en la base de datos. Los resultados para cada métrica de complejidad se presentan en un diagrama de barras, el cual presenta la frecuencia porcentual de los documentos clasificados en cada categoría como se muestra en la Figura 11. En el menú desplegable, ubicado en la parte superior de esta sección (ver Figura 11), se puede seleccionar el año del cual se desean visualizar los resultados; para el año seleccionado, el reporte mostrará los resultados para cada una de las métricas calculadas previamente, las cuales son: cuentas condicionales, la métrica Dale-Chall, el indicador de Shannon y la métrica de Spaulding’s Spanish Readability (SSR).

Figura 10: Fecha de corte y panel de navegación



Fuente: elaboración propia

Figura 11: Medidas de complejidad – visualización inicial

Fuente: elaboración propia

El botón Clasificación del Panel de navegación, dirige a la ventada donde se encuentran los resultados tanto para la clasificación sustancial, la clasificación por sectores y la subclasificación de estos sectores para el año seleccionado. El gráfico de barras para la clasificación Sustancial contiene los porcentajes de textos clasificados como sustanciales y no sustanciales, de la misma manera el gráfico para la clasificación por sectores muestra el porcentaje de textos clasificados en cada uno de los sectores económicos. Para la subclasificación, en la parte superior del gráfico por medio de la lista desplegable, la herramienta permite seleccionar el sector que desea visualizar como se muestra en la Figura 12, enseguida se visualizarán los porcentajes de la cantidad de documentos clasificados en cada subsector del sector seleccionado, en el ejemplo de la Figura 12 se seleccionó el sector "Minería", por lo que el gráfico mostrará los porcentajes obtenidos para los subsectores: *extracción de petróleo y gas, extracción de Uranio y Torio, minerales no metalíferos y minerales metalíferos*.

Figura 12: Gráficos para visualizar los resultados de la subclasificación de cada sector

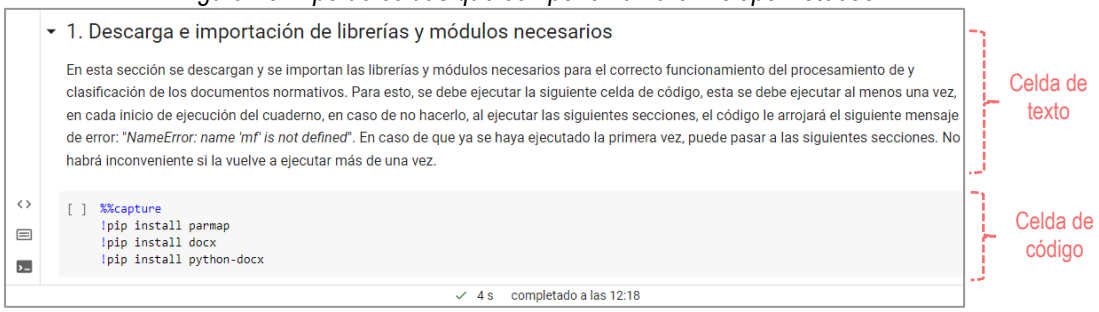
Fuente: elaboración propia

4.7. Herramienta Notebook

Esta herramienta desarrollada en Google Colab, permite realizar todo el proceso por el que deben pasar los documentos regulatorios en 4 simples pasos, para cada uno de los pasos, la herramienta presenta una breve descripción textual y las instrucciones correspondiente. El proceso básicamente consiste en la clasificación de documentos normativos, tanto sustancial como por Sectores Económicos, la clasificación de estos sectores en subsectores, así como el cálculo de las diferentes métricas que permiten estimar el grado de dificultad en la lectura de estos documentos. El proceso completo, que se puede ejecutar en la herramienta, consta de 4 pasos principales (o secciones), estos pasos son:

1. Descarga e importación de librerías y módulos necesarios
2. Carga, procesamiento y clasificación
3. Generación del reporte CSV
4. Generación del reporte HTML

Figura 13: Tipo de celdas que componen un archivo tipo Notebook



Fuente: elaboración propia

Cada sección entonces consta de una celda de texto y de una celda de código como se muestra en la Figura 13, la celda de texto contiene la descripción de las funciones que se realizan allí, y unas breves instrucciones, mientras que la sección de código permite la ejecución de esas funciones. En el ejemplo particular de la Figura 13, se muestra el caso del paso uno, el cual permite descarga e importación de librerías y módulos necesarios.

5. Conclusiones y recomendaciones

El presente documento ha presentado una metodología integral y efectiva para la clasificación y desagregación de documentos normativos en el contexto de 10 sectores económicos. La implementación del nuevo modelo de clasificación con la reducción de dimensionalidad mediante TSVD, la aplicación de la metodología semi-supervisada para la desagregación de los sectores y la optimización de la métrica de complejidad SSR, han demostrado resultados prometedores para una mejor comprensión y gestión de la normativa regulatoria en el país. Estos avances sientan las bases para una mayor eficiencia en la toma de decisiones y en el análisis del marco regulatorio.

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 la continuación del proyecto.

1. La implementación del ajuste del modelo de clasificación de sectores económicos mediante la técnica T-SVD para la reducción de dimensionalidad ha demostrado resultados significativos en términos de precisión y recall. En promedio, se logró una mejora de hasta 4 puntos porcentuales en precisión y 5 puntos porcentuales en recall en comparación con el promedio obtenido en el modelo inicial de esta fase del proyecto. Estos avances representan una mayor confiabilidad en la asignación de las normativas a su respectivo sector económico, lo que, a su vez proporciona una mejor comprensión de la dinámica normativa y contribuye a una toma de decisiones más informada y acertada.
2. La metodología semi-supervisada, basada en la combinación de palabras clave primarias y activadoras, ha demostrado ser una aproximación eficiente y precisa en el proceso de clasificación de los documentos normativos en 43 subsectores económicos. Gracias a esta estrategia, se obtuvieron datos de entrenamiento específicos para cada subsector, los cuales fueron utilizados para entrenar los 8 modelos KNN. Esta combinación de enfoques enriqueció significativamente el proceso de desagregación y subclasificación de los documentos, permitiendo asignar 52.416 documentos previamente clasificados como sustanciales a su respectivo subsector. Estos resultados refuerzan la confiabilidad y efectividad del modelo de clasificación, proporcionando una visión más detallada y completa de la normativa vigente.
3. Se implementó una métrica SSR que permite estimar el grado de dificultad en la lectura de los textos normativos en su lenguaje natural. Gracias a la capacidad de ajustar su fórmula de manera lineal, se ha obtenido un ajuste significativo basado en la realimentación proporcionada por el equipo de Mejora Regulatoria. Este enfoque permitió una clasificación más acertada y realista de la complejidad de los textos normativos, brindando una herramienta más efectiva para analizar y gestionar la información regulatoria en el contexto de Colombia.
4. Los resultados obtenidos al aplicar la métrica SSR revelan que la gran mayoría de los documentos se clasifican en la categoría de *Dificultad Moderada* (49.1% de los documentos). Esta mejora en la clasificación es resultado del ajuste logrado con la ayuda del equipo de Mejora Regulatoria, ya que inicialmente, la mayoría de los documentos se clasificaron en la categoría *Difícil*.
5. Mediante el desarrollo de la herramienta de tipo Notebook, se logró condesar todos los scripts de código consolidados en las diferentes fases del proyecto hasta el momento, lo cual permite que los usuarios de la herramienta puedan ejecutar todo el proceso desde el ingreso de los documentos en PDF, hasta la obtención de la actualización del reporte total, donde se encuentran los resultados obtenidos para cada documento, además de obtener el reporte interactivo para la visualización de estos resultados.

6. Socialización

El proyecto se ha socializado debidamente con el Grupo de Mejora Regulatoria, perteneciente a la Subdirección General de Prospectiva y Desarrollo Nacional, que es precisamente la dependencia del Departamento Nacional de Planeación quien solicitó el proyecto.

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.

ANEXOS

Anexo 1 Subclasificación de sectores económicos, según las divisiones la CIIU

Tabla 7: Clasificación de los sectores económicos según la CIIU, por sector y por subsector

Sector	Subsector	Nombre subsector
1	1	Agricultura, ganadería, caza y actividades de servicios conexas
1	2	Silvicultura, extracción de madera y actividades de servicios conexas
1	3	Pesca, acuicultura y actividades de servicios relacionadas
2	1	Extracción de carbón, carbón lignítico y turba
2	2	Extracción de petróleo crudo y de gas natural, actividades de servicios relacionadas con la extracción de petróleo y de gas, excepto las actividades de prospección
2	3	Extracción de minerales de uranio y de torio
2	4	Extracción de minerales metalíferos
2	5	Explotación de minerales no metálicos
3	1	Elaboración de productos alimenticios y bebidas
3	2	Fabricación
3	3	Confección de prendas de vestir; adobo y teñido de pieles
3	4	Curtido y adobo de cueros; fabricación de calzado; fabricación de artículos de viaje, maletas, bolsos de mano y similares; artículos de talabartería y guarnicionería
3	5	Coquización, fabricación de productos de la refinación del petróleo y combustible nuclear
3	6	Reciclaje
3	7	Transformación de la madera y fabricación de productos de madera y de corcho, excepto muebles; fabricación de artículos de cestería y espartería
3	8	Actividades de edición e impresión y de reproducción de grabaciones
4	1	Suministro de electricidad, gas, vapor y agua caliente
4	2	Captación, depuración y distribución de agua
5	1	Construcción
6	1	Comercio, mantenimiento y reparación de vehículos automotores y motocicletas, sus partes, piezas y accesorios; comercio al por menor de combustibles y lubricantes para vehículos automotores
6	2	Comercio al por mayor y en comisión o por contrata, excepto el comercio de vehículos automotores y motocicletas; mantenimiento y reparación de maquinaria y equipo
6	3	Comercio al por menor, excepto el comercio de vehículos automotores y motocicletas; reparación de efectos personales y enseres domésticos
6	4	Hoteles, restaurantes, bares y similares
7	1	Transporte por vía terrestre; transporte por tuberías
7	2	Transporte por vía acuática
7	3	Transporte por vía aérea
7	4	Actividades complementarias y auxiliares al transporte; actividades de agencias de viajes
7	5	Correo y telecomunicaciones

8	1	Intermediación financiera, excepto el establecimiento y gestión de planes de seguros, de pensiones y cesantías
8	2	Establecimiento y gestión de planes de seguros, de pensiones y cesantías, excepto los planes de seguridad social de afiliación obligatoria
8	3	Actividades de servicios auxiliares de la intermediación financiera
8	4	Actividades inmobiliarias
8	5	Alquiler de maquinaria y equipo sin operarios y de efectos personales y enseres domésticos
8	6	Informática y Actividades conexas
8	7	Investigación y desarrollo
8	8	Otras actividades empresariales
9	1	Administración pública y defensa; planes de seguridad social de afiliación obligatoria
9	2	Educación
9	3	Servicios sociales y de salud
9	4	Otras actividades de servicios comunitarios, sociales y personales
9	5	Actividades de hogares privados como empleadores y actividades no diferenciadas de hogares privados como productores
9	6	Organizaciones y órganos extraterritoriales
10	1	Otros (Administrativos)

Fuente: Elaboración propia

Anexo 2 Diccionario de palabras

Diccionario de palabras primarias y activadoras para cada subsector de cada sector económico, utilizado por el algoritmo de búsqueda (*diccionario_subclasificacion.csv*)