

ASIGNATURA: INTELIGENCIA DE NEGOCIOS 2023-II
PROFESOR: FABIÁN CAMILO PEÑA LOZANO
ESTUDIANTE: JAVIER CERINO, DANIEL ARANGO, MARCO ZULIANI
CÓDIGOS: 202020873, 202110646, 202022412
PROGRAMA DE ESTUDIOS: INGENIERÍA DE SISTEMAS Y COMPUTACIÓN



PROYECTO 1

Tabla de Contenido

1. Entendimiento del negocio y enfoque analítico	2
2. Entendimiento y preparación de los datos.....	3
3. Modelado y evaluación.....	5
3.1. Modelo Random Forest utilizando datos de BoW	5
3.2. Modelo Random Forest utilizando datos de TF-IDF	7
3.3. Modelo Logistic Regression	8
4. Resultados.....	10
5. Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido.....	10
6. Trabajo en equipo.....	11
7. Referencias.....	12

1. Entendimiento del negocio y enfoque analítico

El contexto del problema busca identificar la pertenencia de unos artículos dados en los diferentes objetivos de desarrollo sostenible (ODS¹) que utiliza la ONU². Con este proyecto se busca desarrollar un modelo de clasificación, con técnicas de aprendizaje automático, que permita relacionar de manera automática un texto según los ODS. Nuestro grupo recibió la asignación de los tópicos de: agua limpia y saneamiento (número 6 de los ODS), energía asequible y no contaminante (número 7 de los ODS) y paz, justicia e instituciones sólidas (número 16 de los ODS).

Oportunidad/problema Negocio	La oportunidad de negocio que motiva la creación de este proyecto es la necesidad de clasificar un artículo, dado su contenido, en las categorías de los objetivos de desarrollo sostenible instaurados por la ONU.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar	Para poder satisfacer la oportunidad de negocio se debe implementar una función que permita transformar los textos a una estructura de datos que permita el entrenamiento de un modelo. Para esto decidimos hacer uso de 2 tipos de algoritmos para la analítica de textos. Estos algoritmos son el <i>Bag of Words</i> y el <i>TF-IDF</i> . Para los modelos de aprendizaje automático que se van a construir se propone el uso de los algoritmos <i>Random Forest</i> y <i>Logistic Regression</i> .
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización que se beneficiaría directamente con la oportunidad del negocio definida es la ONU. El área que utilizará los modelos de clasificación contruidos es la UNFPA ³ . Por lo anterior, la ONU podría ver una reducción en los gastos y recursos invertidos en la UNFPA mientras que este departamento puede realizar un análisis más eficiente para poder tomar las medidas y/o hacer los controles necesarios en las áreas respectivas de los ODS para que se identifiquen las necesidades y se planteen las soluciones requeridas según los artículos.

¹ (Moran, s. f.)

² (Nations, s. f.)

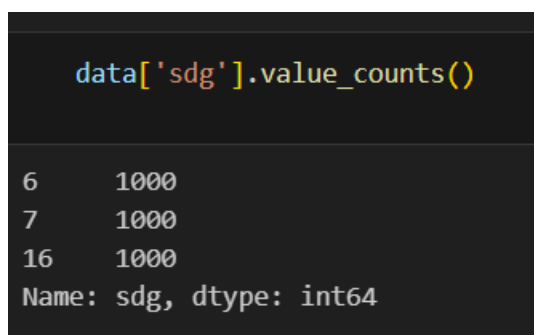
³ (UNFPA Colombia | UNFPA en Colombia, s. f.)

Contacto con experto externo al proyecto	Maria Londoño Tejada, m.londonot2@uniandes.edu.co Reunión: miércoles 18 - 6pm Canal: Zoom
---	---

Tabla 1

2. Entendimiento y preparación de los datos.

Para el desarrollo de este proyecto se nos proveyó de un archivo de Excel con un total de 3000 registros. Cada uno de estos registros contaban con un texto escrito en español y con su respectiva etiqueta según la clasificación de los ODS. Como se mencionó en la primera parte nuestro grupo estará encargado de solo 3 de los grupos ya mencionados, el 6, 7 y 16. Para la realización de este análisis y de las correcciones requeridas se hará uso del lenguaje de programación *Python* junto con diversas librerías. Para la carga de datos se optó por usar la estructura de datos *Dataframe*, implementada por la librería de *Pandas*. Como se puede ver en la imagen 1, la distribución de los datos es perfectamente equitativa teniendo una distribución de 1000 registros para cada uno de los objetivos a analizar.



```
data['sdg'].value_counts()

6      1000
7      1000
16     1000
Name: sdg, dtype: int64
```

Imagen 1.

Este detalle es extremadamente importante ya que no permite poder contar con una mejor fuente de datos para el entrenamiento del modelo evitando así un caso de sub-ajuste.

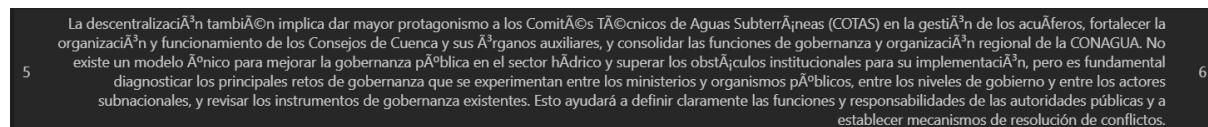
El primer paso que se realizó al momento del entendimiento de los datos es cargar las palabras vacías⁴, también conocidas como stop words en inglés. Estas palabras vacías son aquellas palabras que no deben ser consideradas al momento de indexar las palabras para realizar consultas, la mayor parte de estas palabras son pronombres, adverbios, conjunciones entre otras. Esta lista de palabras se puede obviar debido a que, al momento de hablar español, si son indispensables para una comunicación efectiva, pero al momento del análisis de textos son ruido debido a su uso tan estandarizado.

Una vez cargados los datos al notebook de Jupyter, observamos que el documento en Excel contaba con dos columnas, la llamada 'Textos_espanol' contiene toda la información de los textos que se van a utilizar. Por otro lado, la columna, 'sdg' cuenta con los índices de los ODS en los que se clasificaron cada uno de dichos artículos. El segundo artículo que se nos entregó es otro documento de Excel en el que se pueden

⁴ (K, 2014)

observar un total de 980 registros, la diferencia más notoria es que este documento, no cuenta con ninguna etiqueta que asocie los textos a una categoría de los objetivos de desarrollo sostenible de la ONU.

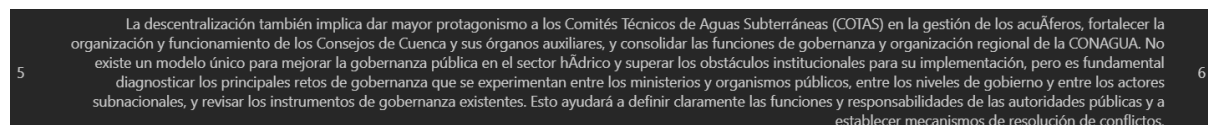
Al revisar con más detalle el documento de datos que se nos fue entregado, se identificaron que ciertos escritos cuentan con caracteres anómalos que se generaron debido a la presencia de caracteres especiales tales como tildes o acentos circunflejos. En la imagen 2 se muestra un ejemplo de dichas entradas anómalas identificadas.



La descentralización también implica dar mayor protagonismo a los Comités Técnicos de Aguas Subterráneas (COTAS) en la gestión de los acuíferos, fortalecer la organización y funcionamiento de los Consejos de Cuenca y sus Órganos auxiliares, y consolidar las funciones de gobernanza y organización regional de la CONAGUA. No existe un modelo único para mejorar la gobernanza pública en el sector hídrico y superar los obstáculos institucionales para su implementación, pero es fundamental diagnosticar los principales retos de gobernanza que se experimentan entre los ministerios y organismos públicos, entre los niveles de gobierno y entre los actores subnacionales, y revisar los instrumentos de gobernanza existentes. Esto ayudará a definir claramente las funciones y responsabilidades de las autoridades públicas y a establecer mecanismos de resolución de conflictos.

Imagen 2

Para solucionar este problema decidimos utilizar la librería *ftfy* en particular la función *fix_text* que permite transformar estos caracteres en sus respectivas contrapartes antes de que se corrompieran. Una vez aplicada esta transformación obtenemos que estas frases ya no se encuentran con valores atípicos, ver imagen 3.



La descentralización también implica dar mayor protagonismo a los Comités Técnicos de Aguas Subterráneas (COTAS) en la gestión de los acuíferos, fortalecer la organización y funcionamiento de los Consejos de Cuenca y sus órganos auxiliares, y consolidar las funciones de gobernanza y organización regional de la CONAGUA. No existe un modelo único para mejorar la gobernanza pública en el sector hídrico y superar los obstáculos institucionales para su implementación, pero es fundamental diagnosticar los principales retos de gobernanza que se experimentan entre los ministerios y organismos públicos, entre los niveles de gobierno y entre los actores subnacionales, y revisar los instrumentos de gobernanza existentes. Esto ayudará a definir claramente las funciones y responsabilidades de las autoridades públicas y a establecer mecanismos de resolución de conflictos.

Imagen 3

El siguiente paso que se realizó fue la transformación de todas las palabras encontradas a letras minúsculas, funcionaría exactamente igual si se pasaran todas a mayúsculas, no obstante, por facilidad lectora se procedió como mencionado. Con esta última transformación logramos asegurar que todas las palabras iguales se asociaran por significado y no se tuvieran registros del tipo, “Perro” y “perro” que se comportaran como dos instancias diferentes.

A continuación, en nuestra preparación de datos se removieron los signos de puntuación esto es debido a que, muy similar a lo que ocurre con la transformación de las mayúsculas y minúsculas, los algoritmos que se usaran más adelante consideran palabras como “perro.” y “perro” como dos instancias diferentes en donde la única diferencia era dicho carácter no alfabético. Para evitar este tipo confusiones y similares se procedió a eliminar la fuente del problema.

Para finalizar el entendimiento y limpieza de datos se procede a realizar un último cambio y es transformar todos los números escritos con caracteres numéricos (ej. 1, 2, 3, etc.) a escritura alfabética (ej. uno, dos, tres, etc.) esto para que los algoritmos de clasificación no encuentren problemas al momento de construir el modelo.

Para la preparación de datos previa a la construcción del modelo se utilizaron dos técnicas de analítica de textos para transformar todos los escritos en un dataframe adecuado para ello. Las dos técnicas utilizadas fueron los algoritmos de *Bag of Words* y TF-IDF (term frequency, inverse document frequency). El primero de estos tiene un comportamiento en el que se busca contar la cantidad de apariciones de cada una

de las palabras en los textos analizados, sucesivamente se registra en el dataframe la cantidad de veces que aparece cada una de las palabras en cada texto. El segundo algoritmo busca crear una medición no solo basada en la frecuencia en que cada palabra aparece en todos los textos analizados si no que se realiza una media numérica que expresa cuán relevante es dicha palabra con respecto a la colección total de textos analizados, para la construcción de este algoritmo se tiene que contabilizar la cantidad de veces que aparece la palabra en el texto a la vez que cuantas veces aparece esa palabra en todos los textos analizados⁵. Estos dos algoritmos vienen implementados por la librería de *sklearn*.

Cabe aclarar que para poder la construcción de cualquiera de estos modelos es necesario que los datos que se le vayan a pasar no contengan las palabras vacías del idioma que se esté evaluando. En nuestro caso para el entrenamiento se realizó esta eliminación de forma manual, no obstante, las funciones de *sklearn* permiten que se les ingrese una lista de estas palabras para que las ignore al momento de la construcción del modelo.

3. Modelado y evaluación.

El primer paso que se realizó para poder llevar a cabo el entrenamiento de los modelos fue la división de los datos en un conjunto para entrenamiento y uno para testeo. Para tal fin, se utilizó la función *train_test_split* de la librería *sklearn* sobre la columna “Textos_espanol”, estableciendo como variable objetivo la columna “sdg” del dataframe y como 0.3 el porcentaje de datos para el conjunto de testeo. Además, se utiliza el parámetro *stratify* sobre la columna “sdg” para mantener la distribución equitativa de los valores de la variable objetivo tanto en los datos de entrenamiento como en los de testeo, en este caso de 1/3 de los datos.

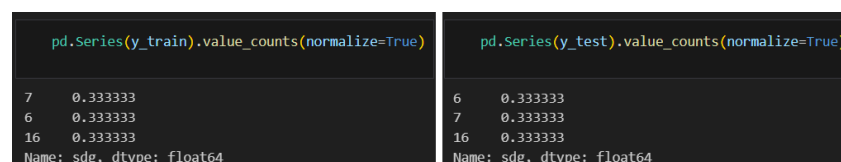


Imagen 4

Posteriormente, se utilizan los dataframes creados anteriormente con las técnicas de analítica de textos para el entrenamiento de tres modelos. Para la clasificación de los textos, se utilizó el algoritmo de *Random Forest* para los primeros dos y el de *Logistic Regresion* para el último.

3.1. Modelo Random Forest utilizando datos de *BoW* (Realizado por: **Marco Zuliani**)

Posterior al entrenamiento de este primer modelo, se analizan las palabras de mayor importancia en el dataframe. Como se puede ver a continuación en la imagen 5, la palabra “violencia” es la de mayor relevancia con una presencia de poco más del 5%, seguida por la palabra “valer” con poco más de un 4%.

⁵ (Hamdaoui, 2021)

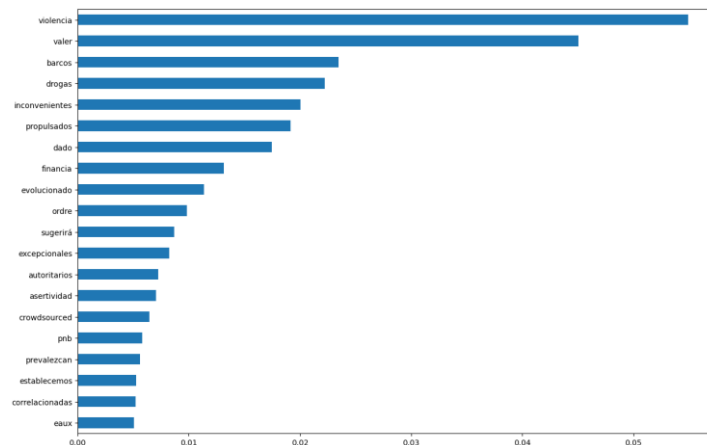


Imagen 5

Al analizar el modelo generado, se evidencia que el número de árboles (estimadores) es de 100 y que la media de las profundidades de estos es de 95.62.

```
Number of trees: 100
Trees depth (mean): 95.62
```

Imagen 6

Luego, se utiliza el modelo para predecir la clasificación de los textos de entrenamiento y de testeo y determinar la efectividad del algoritmo en la realización de esta tarea. Realizando un análisis de la matriz de confusión de la predicción sobre los datos de entrenamiento (Imagen 6^a), se puede notar al observar la diagonal de la matriz, como la predicción se realiza de manera perfecta y no se presentan falsos positivos, falsos negativos ni verdaderos negativos. Por otro lado, al analizar la matriz de confusión de la predicción sobre los datos de testeo (Imagen 6^b), es evidente como hay una muy buena distribución de verdaderos positivos, pero encontramos ciertos valores de falsos negativos, falsos positivos y verdaderos negativos.

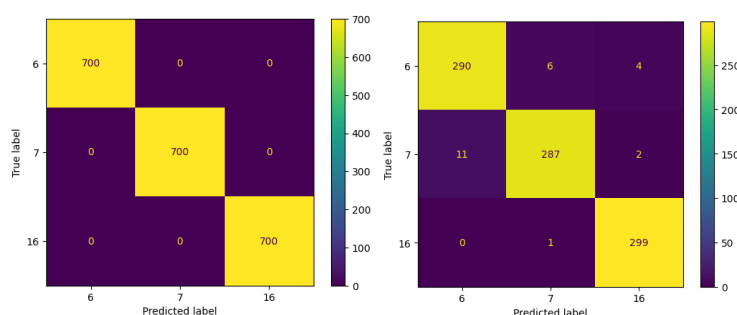


Imagen 7^a

Imagen 7^b

Adicionalmente, para ambos casos se utiliza la "Precision", el "Recall" y el "F1-score" para poder realizar un análisis más certero de la predicción del modelo. A continuación, se muestran los valores obtenidos.

Precision: 1.0 Recall: 1.0 F1: 1.0	Precision: 0.973324498181532 Recall: 0.9733333333333333 F1: 0.9732726515563201
Imagen 8 ^a	Imagen 8 ^b

Es evidente que, en la predicción de los datos de entrenamiento, todos los textos fueron clasificados de manera correcta por lo que el valor obtenido en todas las métricas es de 1 (Imagen 8^a). De manera similar, en la predicción de los datos de entrenamiento las métricas son realmente altas, lo que indica que el modelo predice de manera correcta una gran mayoría de los textos previamente desconocidos (Imagen 8^b).

3.2. Modelo Random Forest utilizando datos de TF-IDF (Realizado por: **Javier Cerino**)

Posterior al entrenamiento del segundo modelo, se analizan las palabras de mayor importancia en el dataframe. Como se puede ver a continuación en la imagen 9, la palabra “violencia” es la de mayor relevancia con una presencia de poco más del 5%, seguida por la palabra “valer” con casi un 3%.

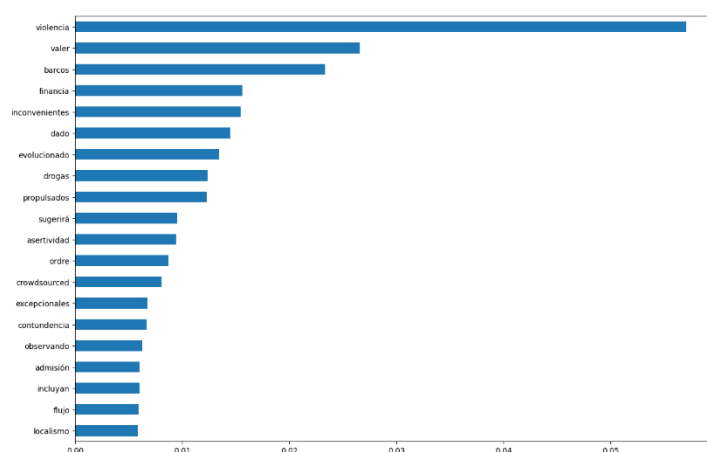


Imagen 9

Al analizar el modelo generado, se evidencia que el número de árboles (estimadores) es de 100 y que la media de las profundidades de estos es de 90.51. Esto nos permite observar como este modelo se ajusta un poco mejor a los datos.

```
Number of trees: 100
Trees depth (mean): 90.51
```

Imagen 10

Luego, se utiliza el modelo para predecir la clasificación de los textos de entrenamiento y de testeo y determinar la efectividad del algoritmo en la realización de esta tarea. Realizando un análisis de la matriz de confusión de la predicción sobre los datos de entrenamiento (Imagen 10^a), se puede notar al observar la diagonal de la matriz, como la predicción se realiza de manera perfecta y no se presentan falsos

positivos, falsos negativos ni verdaderos negativos. Por otro lado, al analizar la matriz de confusión de la predicción sobre los datos de testeo (Imagen 10^b), es evidente como hay una muy buena distribución de verdaderos positivos, pero encontramos ciertos valores de falsos negativos, falsos positivos y verdaderos negativos.

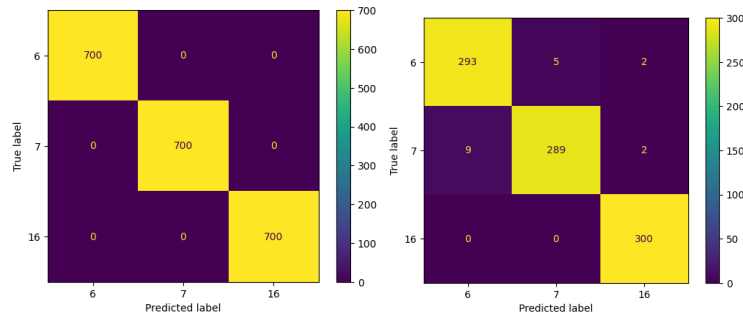


Imagen 11^a

Imagen 11^b

Adicionalmente, para ambos casos se utiliza la "Precision", el "Recall" y el "F1-score" para poder realizar un análisis más certero de la predicción del modelo. A continuación, se muestran los valores obtenidos.

```
Precision: 1.0
Recall: 1.0
F1: 1.0
```

Imagen 12^a

```
Precision: 0.9800113260129194
Recall: 0.98
F1: 0.9799544611393269
```

Imagen 12^b

Es evidente que, en la predicción de los datos de entrenamiento, todos los textos fueron clasificados de manera correcta por lo que el valor obtenido en todas las métricas es de 1 (Imagen 11^a). De manera similar, en la predicción de los datos de entrenamiento las métricas son realmente altas, lo que indica que el modelo predice de manera correcta una gran mayoría de los textos previamente desconocidos (Imagen 11^b).

Realizando una comparación de este modelo con respecto al anterior (Modelo Random Forest utilizando datos de TF-IDF), se puede notar como hubo una mejora de un total de 1% lo que indica que es una mejor aproximación. Esto ya se podía inferir de forma imparcial con la imagen 9 en donde vimos que aumento la relevancia de la palabra "violencia" aumentado así su importancia al momento de predecir el modelo, situación que se logró gracias a la media aritmética y no solo por la cantidad de apariciones de estos términos en el diccionario.

3.3. *Modelo Logistic Regression* (Realizado por: **Daniel Arango**)

Luego de haber entrenado los anteriores modelos con distintos hiperparametros se decidió probar el algoritmo de Logistic Regression que "Es un algoritmo muy utilizado para la clasificación en la industria." (Subasi, 2020), lo cual está muy ligado a su simplicidad y eficiencia con los problemas de clasificación de tipo lineal. Usualmente este modelo estadístico busca clasificar de manera binaria y determinar si pertenece

a una categoría o a otra. No obstante, es posible generalizarlo para poder clasificar con múltiples clases utilizando el algoritmo “One Vs Rest” (OVR).

Como su nombre sugiere, este algoritmo implica seleccionar una clase como objetivo y considerar todas las demás clases como una segunda categoría virtual. Luego, se aplica la regresión logística binaria a esta configuración. Repetimos este proceso para cada una de las clases presentes en el conjunto de datos. Como resultado, se obtiene clasificadores binarios individualizados, cada uno diseñado para identificar una clase específica dentro del conjunto de datos.

```
Precision: 0.9957122303331173
Recall: 0.9957142857142857
F1: 0.9957122419825043
```

Imagen 13^a

```
Precision: 0.9845354622503447
Recall: 0.9844444444444443
F1: 0.9844193572697767
```

Imagen 13^b

Luego de haber realizado el entrenamiento y la predicción con los datos de entrenamiento vectorizados se obtienen las métricas mostradas en las (Imagen 13^a) las cuales nos describen la eficiencia del modelo para el conjunto de datos de entrenamiento, donde vemos que hay una pequeña diferencia entre los otros modelos. Por otro lado, en la imagen (Imagen 13^b) podemos ver como se obtiene un gran porcentaje de precisión y de “recall” el cual nos muestra la eficiencia del algoritmo en identificar correctamente el tipo de ODS al que se refiere el párrafo. Esto a su vez podemos verlo en la matriz de confusión de las imágenes 14a y 14b donde para cada clase es evidente como hay una muy buena distribución de verdaderos positivos, pero encontramos ciertos valores de falsos negativos, falsos positivos y verdaderos negativos.

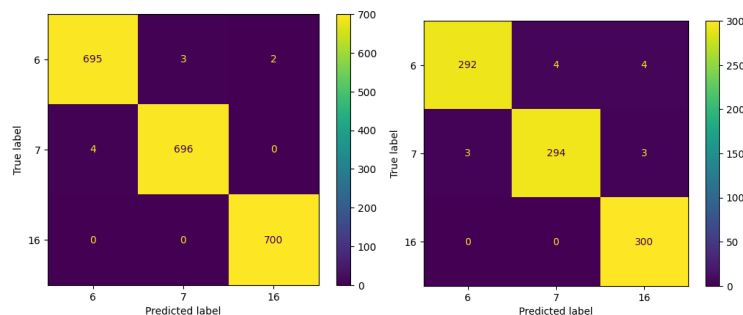


Imagen 14^a

Imagen 14^b

Finalmente, realizando una comparación de este modelo con respecto al anterior (Modelo Random Forest utilizando datos de TF-IDF), se puede notar como hubo una mejora de un total de 0.04% lo que indica que es una mejor aproximación, a pesar de haber disminuido en el puntaje de Train obtenido. Dado que, un modelo que alcanza una puntuación perfecta en el entrenamiento podría estar sobreajustado a los datos de entrenamiento y no ser capaz de lidiar efectivamente con nuevos datos. En cambio, un modelo que logra un rendimiento más equilibrado demuestra que no se está ajustando en exceso y tiene una mayor probabilidad de funcionar bien en situaciones del mundo real, lo que aumenta la confianza en su capacidad para adaptarse a datos futuros.

4. Resultados.

Métrica	BoW Random Forest	TF-IDF Random Forest	TF-IDF Logistic Regression
Precision	0.973324498181532	0.9800113260129194	0.9845354622503447
Recall	0.9733333333333333	0.98	0.9844444444444443
F1-score	0.9732726515563201	0.9799544611393269	0.9844193572697767

Tabla 2

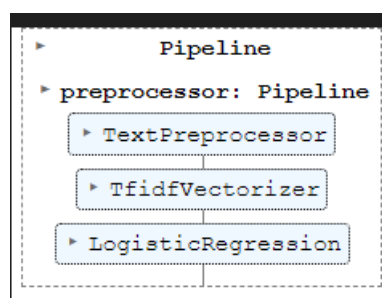


Imagen 15

Como se puede observar en la Tabla 2, al realizar la comparativa entre los 3 modelos construidos el que da una mejor medición en todas las métricas es el Logistic Regression con TF-IDF da un mejor resultado por lo que se seleccionó este modelo para entrenar el pipeline para etiquetar los datos finales. En la imagen 15 podemos observar que contiene el pipeline construido. Por otro lado, podemos observar un ejemplo de 3 textos que fueron etiquetados al ejecutar el pipeline diseñado.

	Textos_espanol	sdg
0	1. 1. Introducción: Las Estructuras del Derecho Penal 2. El Estándar de la Persona Razonable en el Derecho Penal 3. La Responsabilidad Penal de los Delincuentes Suerte resultante y responsabilidad penal 4. 4. Criminalización del sadomasoquismo: negación de lo erótico, instanciación de la violencia 5. Constitucionalismo y límites del Derecho Penal 6. Delincuencia internacional: contexto y contraste Forma jurídica y juicio moral: Eutanasia y suicidio asistido 8. Derecho anormal: La teratología como lógica de criminalización 9. Tensiones de la criminalización: Desierto Empírico, Cambio de Normas y Reforma de la Violación 10. Delitos de Preparación, Intereses de Seguridad y Libertad Política Delitos de preparación, intereses de seguridad y libertad política	16
1	Las aguas subterráneas se han debatido en el contexto de la tarificación y la financiación (OCDE, 2009a y 2009b), la energía (OCDE, 2012b), la gestión de riesgos (OCDE, 2013e) y perspectivas más amplias que abarcan el cambio climático (OCDE, 2013d y 2014a). Las aguas subterráneas también aparecen en las revisiones de las reformas del agua a nivel nacional (por ejemplo, Fuentes, 2011, OCDE, 2013b). Todos estos informes incluyen secciones, subsecciones, párrafos o ilustraciones que se refieren a las aguas subterráneas, pero no transmiten conclusiones políticas específicamente orientadas a los gestores de tipos concretos de aguas subterráneas, sobre todo en el contexto de la agricultura. En primer lugar, una observación coherente es que las aguas subterráneas están generalmente poco estudiadas y que es necesario realizar una evaluación más profunda de las reservas, el uso y las prácticas de gestión de las aguas subterráneas.	6
3	Sin embargo, este crédito fiscal expira en 2012. Wernau (2011a) calcula que hasta una quinta parte de la capacidad de generación de energía del Estado podrá decidir abandonar el mercado en lugar de invertir en las mejoras necesarias, lo que podrá aumentar el precio del carbón en un 65% y ofrecer oportunidades a las fuentes de energía renovables. Ahorra aproximadamente 47 000 toneladas de CO2 al año y sustituye anualmente más de ocho millones de desplazamientos en vehículo privado en la ciudad de Calgary.	7

Imagen 16

5. Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido.

Rol	Tipo de actor	Beneficio	Riesgo
Analista de textos de UNFPA	Usuario – Cliente	Disminuye el tiempo necesario para clasificar un texto para poder realizar reportes de estado de las diferentes áreas de interés para identificar un potencial proyecto a desarrollar.	Una mala predicción de los contenidos de los artículos lo que generaría un impacto grave en la organización de los trabajos malgastando tiempo y recursos de áreas a los que no les compete la información de dicho artículo.

Trabajador de la ONU	Financiador	Una disminución en costos y tiempo que permite analizar más textos en el mismo tiempo permitiendo crear proyectos con una mejor perspectiva.	En caso de que el modelo no funcione la inversión de capital que se realizó en la construcción en este es un desperdicio que se pudo haber utilizado en contratar más empleados capaces de suplir el modelo.
Comunidad publicadora del artículo.	Beneficiado	Permite obtener una retroalimentación y/o apoyo de la ONG o algún ente competente de forma mucho más rápida y eficiente.	Una mayor demora en el momento de recibir financiación o apoyo de expertos debido a la ausencia de categorización de los proyectos que se podrían generar por una mala clasificación de los artículos.

6. Trabajo en equipo.

Para el trabajo en grupo del proyecto 1 se definieron los siguientes roles: **Daniel Arango**: Líder de analítica. **Javier Cerino**: Líder de negocio. **Marco Zuliani**: Líder de proyecto. Debido a que somos 3 integrantes todos vamos a asumir el rol de Líder de datos ya que todos trabajamos en conjunto para que los datos pudieran ser utilizados. Se realizaron las siguientes reuniones: martes 3 de octubre: reunión de lanzamiento y planeación. En esta se leyó el enunciado y se definieron los roles a usar. De igual forma se planearon las siguientes reuniones del equipo. Viernes 6 de octubre: reunión de ideación: En esta se definieron los temas de organización/ empresa/ institución para darles el apoyo necesario con la construcción de nuestro modelo. La primera reunión de seguimiento se realizó el martes 10 de octubre en donde se definieron a profundidad los modelos a desarrollar y se empezó el estudio de cada estudiante para cada uno de sus modelos desarrollados. El viernes 13 de octubre se concluyó la limpieza de datos y la construcción de 2 modelos de 3. El sábado 14 se redactó el documento Word a la vez que se construía el tercer modelo. El domingo 15 se eligió el mejor modelo y se etiquetaron los datos a entrega, se grabó el video y se creó el repositorio para la entrega final.

7. Referencias

- Hamdaoui, Y. (2021, Marzo 24). *TF(Term Frequency)-IDF(Inverse Document Frequency) from scratch in python* . Medium. <https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558>
- K, G. L. (2014, octubre 5). Lista de stop words o palabras vacías en español. *SEO para Google*. <https://googleseo.marketing/lista-de-stop-words-o-palabras-vacias-en-espanol/>
- Moran, M. (s. f.). La Agenda para el Desarrollo Sostenible. *Desarrollo Sostenible*. Recuperado 14 de octubre de 2023, de <https://www.un.org/sustainabledevelopment/es/development-agenda/>
- Nations, U. (s. f.). *Naciones Unidas | Paz, dignidad e igualdad en un planeta sano*. United Nations; United Nations. Recuperado 14 de octubre de 2023, de <https://www.un.org/es/>
- UNFPA Colombia | UNFPA en Colombia*. (s. f.). Recuperado 14 de octubre de 2023, de <https://colombia.unfpa.org/es/unfpa-en-colombia>
- Subasi, A. (2020). Machine learning techniques. En A. Subasi (Ed.), *Practical Machine Learning for Data Analysis Using Python* (pp. 91–202). Elsevier.
- Multiclass logistic regression Using sklearn*. (2020, junio 1). Kaggle.com; Kaggle. <https://www.kaggle.com/code/satishgunjal/multiclass-logistic-regression-using-sklearn>