Departamento de Ingeniería de Sistemas y Computación

Inteligencia de Negocios

Proyecto 1: Analítica de textos

Juan Sebastián Pinzón Sampedro – 202013144

Carlos Andrés García Gómez – 202013993

Andres Felipe Garcia Bernal – 201716813

Contenido

[Parte 1: Entendimiento del negocio y enfoque analítico 1](#_Toc148283379)

[Parte 2: Entendimiento y preparación de datos 2](#_Toc148283380)

[Parte 3: Modelado y evaluación 3](#_Toc148283381)

[Parte 4: Resultados 4](#_Toc148283382)

[Parte 5: Mapa de actores 4](#_Toc148283383)

[Parte 6: Trabajo en equipo 5](#_Toc148283384)

# Parte 1: Entendimiento del negocio y enfoque analítico

* **Oportunidad/Problema de negocio:**

Uno de los procesos que requiere de un mayor esfuerzo es la clasificación de la información textual que es recopilada, ya que es una tarea que consume gran cantidad recursos y para la cual se requiere un experto.

* **Enfoque analítico:**

Particularmente, el requerimiento solicitado por la organización es el de clasificar un conjunto de datos, en este caso datos de texto (oraciones), en alguno de los Objetivos de desarrollo sostenible (ODS), que sea más afín al texto. Específicamente se calificarán los datos en 3 diferentes ODS, el 6,7 o 16. Para ello se implementarán algoritmos de aprendizaje automático para clasificación. Se propone hacer uso de los algoritmos de Random forest, Logistic regression y decisión Trees dado que son algoritmos que funcionan bien para el caso de clasificación de textos. Por otro se hará uso de otros algoritmos para la limpieza de los datos de modo que los algoritmos de clasificación anteriormente mencionados puedan funcionar.

* **Organización y rol dentro de ella que se beneficia con la oportunidad definida:**

La organización que se beneficia es el Fondo de Poblaciones de las Naciones Unidas (UNFPA), particularmente el área de recursos humanos y el área de analítica de datos, debido a que desean implementar diferentes estrategias de clasificación de textos para automatizar el análisis de las opiniones de habitantes locales que son recogidas mediante sus diferentes herramientas de participación ciudadana.

* **Contacto con experto externo al proyecto:**

****

Ilustración 1 Listado de ODS tomado de: <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>

# Parte 2: Entendimiento y preparación de datos

Para el entendimiento de los datos se tuvo en cuenta la forma en la que esta presentada la información de la organización. En esta, se tiene unos archivos de Excel el cual se organizan en dos columnas: “Texto\_espanol” y “sdg”. En la columna “Texto\_espanol” se registra lo dicho por un ciudadano con respecto a su preocupación con respecto a un objetivo de desarrollo sostenible y en la columna “sdg”, se registra el número de este ODS.

Subsecuentemente, dado que es posible que existan errores de tipado por parte de los usuarios, se realiza la debida limpieza de la información para poder manejarla con mayor eficiencia. En este proceso de limpieza se convirtieron valores de mayúsculas a minúsculas, eliminación de etiquetas HTML, eliminación de puntuación, eliminación de números, normalización de espacios en blanco y eliminación de caracteres especiales. De esta manera, se podrá agrupar mejor la información y realizar un análisis mas exacto de las palabras que permiten determinar a qué ODS hace referencia un usuario con solo conocer lo que escribe.

Posteriormente, se remueven las conocidas Stop Words que son palabras generales como conectores que de por sí solas, no aportan información relevante para el análisis y casi siempre generan ruido en el mismo, obstruyendo así el proceso de la creación de un modelo confiable que permita determinar el valor de un ODS de acuerdo a un texto.

Luego se aplica sobre el conjunto de datos la lematización para reducir aún mas el conjunto de palabras. Eso se hace modificando a todas las palabras a su forma base o raíz para que, de esta manera, no hayan demasiados conjuntos que representen lo mismo. Esto permite facilitar los procesos de análisis que se efectuaran posterior a la ejecución del modelo.

Finalmente, se almacenará la información que fue “limpiada” en una nueva tabla para su posterior utilización.

# Parte 3: Modelado y evaluación

Para esta parte se decidió implementar los siguientes algoritmos de clasificación:

* Logistic Regression (implementado por: Sebastián Pinzón)
* Random Forest (implementado por: Carlos Garcia)
* Decision Trees (implementado por: Andres Garcia)

Los métodos cuantitativos de evaluación de los modelos que se utilizaron fueron:

* Matriz de confusión
* F1-score
* Precision
* Recall
* Curva ROC y área bajo la curva ROC (AUC-ROC)

**Decision Trees**

Los árboles de decisión son un algoritmo de aprendizaje automático supervisado utilizado para tareas de clasificación y regresión el cual funciona de manera iterativa. El funcionamiento de este algoritmo consiste en seleccionar un atributo de decisión para posteriormente dividir su registro en conjuntos de datos. Estos dos procesos de hacen de manera iterativa un número de veces hasta hallar una condición de parada, es decir, cuando todos los problemas de un subconjunto pertenecen al mismo conjunto, se alcanza la profundidad máxima en el árbx`ol, se alcanza el número mínimo de elementos en sus hojas u otra condición brindada por el negocio.

Para hacer una predicción con un árbol de decisión, se comienza desde el nodo raíz y se sigue el camino correspondiente a través del árbol siguiendo las decisiones basadas en los valores de las características del dato de entrada. Cuando se llega a un nodo hoja, se devuelve la clase (en clasificación) o el valor objetivo (en regresión) asociado a ese nodo hoja como la predicción.

Finalmente, aunque el algoritmo es útil en parte porque los resultados son fácilmente interpretables, es común que estos algoritmos tengan alto overfitting, puesto que, si se tienen demasiadas iteraciones dentro del árbol es posible que el modelo se ajuste demasiado a los datos de entrenamiento y termine prediciendo nuevos datos de manera errónea.

**Random Forest**

El algoritmo Random Forest es una técnica que se utiliza en el aprendizaje automático para la clasificación y regresión. Funciona creando múltiples árboles de decisión y combinándolos para obtener predicciones más precisas y estables. Para el caso de nuestro interés, es decir, clasificación, primero el algoritmo crea los árboles de decisión y posteriormente se realiza la “votación”. Cada árbol "vota" por la clase a la que cree que pertenece el punto de datos, y la clase que recibe la mayoría de los votos se elige como la predicción final.

**Logistic Regression**

La regresión logística es un algoritmo de aprendizaje automático supervisado utilizado para problemas de clasificación binaria o multiclase. A diferencia de los algoritmos de regresión lineal que se utilizan para predecir valores continuos, la regresión logística se emplea para predecir la probabilidad de que una instancia pertenezca a una de las clases en un problema de clasificación.

La primera diferencia entre estas dos regresiones es el tipo de curva. En una regresión linear la curva se calcula con los cuadrados de los residuos, ademas como indica su nombre es de carácter lineal. En la regresión lineal se usa una medida llamada máxima verosimilitud, esta se calcula creando una curva logística que tiene forma de S y variándolo hasta que coloque los puntos con la mayor probabilidad de que sean calificados en la categoría correcta.

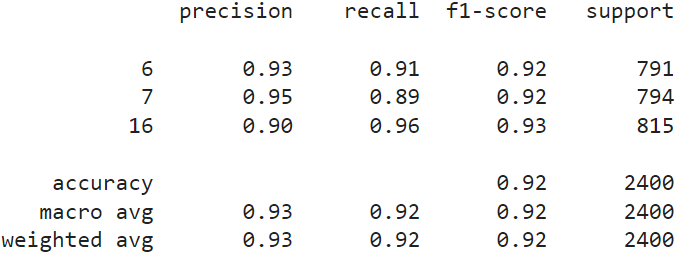
# Parte 4: Resultados

Para el análisis de los resultados, se tuvo en cuenta las variables resultantes de “precisión”, “recall” y “F1” donde cada una es más o menos relevante dependiendo de la necesidad de la organización.

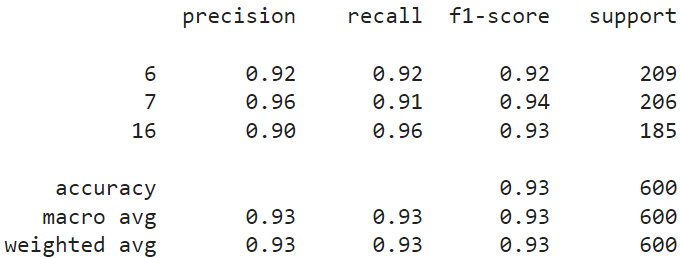
Los resultados que se presentaron son los siguientes.

**Logistic Regression**

En el algoritmo de regresión logística se obtuvieron los siguientes valores en entrenamiento y en prueba.



Resultado de la predicción con valores de entrenamiento.

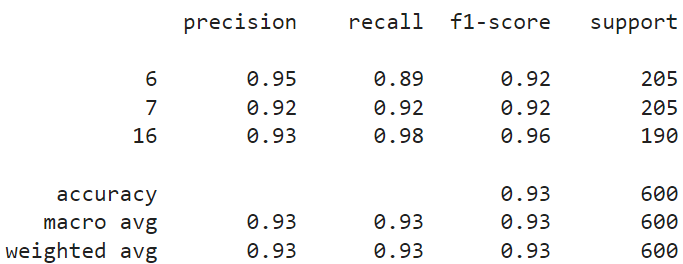


Resultado de la predicción con valores de prueba.

De acuerdo con esta información se puede concluir que no existe casi overffiting, por lo cual es modelo de entrenamiento es muy similar al de prueba. Además, dado que hay un valor de F1 tan alto, se puede afirmar que el modelo es bastante confiable para predecir los valores de ODS. Por otro lado es importante denotar que existen pequeños cambios en la precisión y el recall dependiendo del tipo de ODS que se este analizando (7 y 16) por lo cual, si al negocio no le molesta este pequeño margen de error que existe para determinar un tipo de ODS, la regresión logística puede ser el modelo preferido a utilizar para realizar estas predicciones.

**Random forest**

En el algoritmo de Random forest logística se obtuvieron los siguientes valores.



Resultado de la predicción con los valores de prueba

Dado que se utilizan ciertos registros de la tabla para entrenar el modelo, no tiene sentido hacer la predicción con esos datos dado que tendría 100% de precision, recall y f1-score.

Sabiendo eso, podemos afirmar que el algoritmo de Random Forest posee un wighted avg igual al algoritmo de regresión logística, sin embargo, también existen valores que fluctúan como el recall de 6 por lo que nuevamente, depende del negocio el aceptar ese porcentaje de riesgo de un falso negativo o no.

**Decision Trees**

En el algoritmo de árboles de decisión se obtuvieron los siguientes valores.

A number of numbers in a row

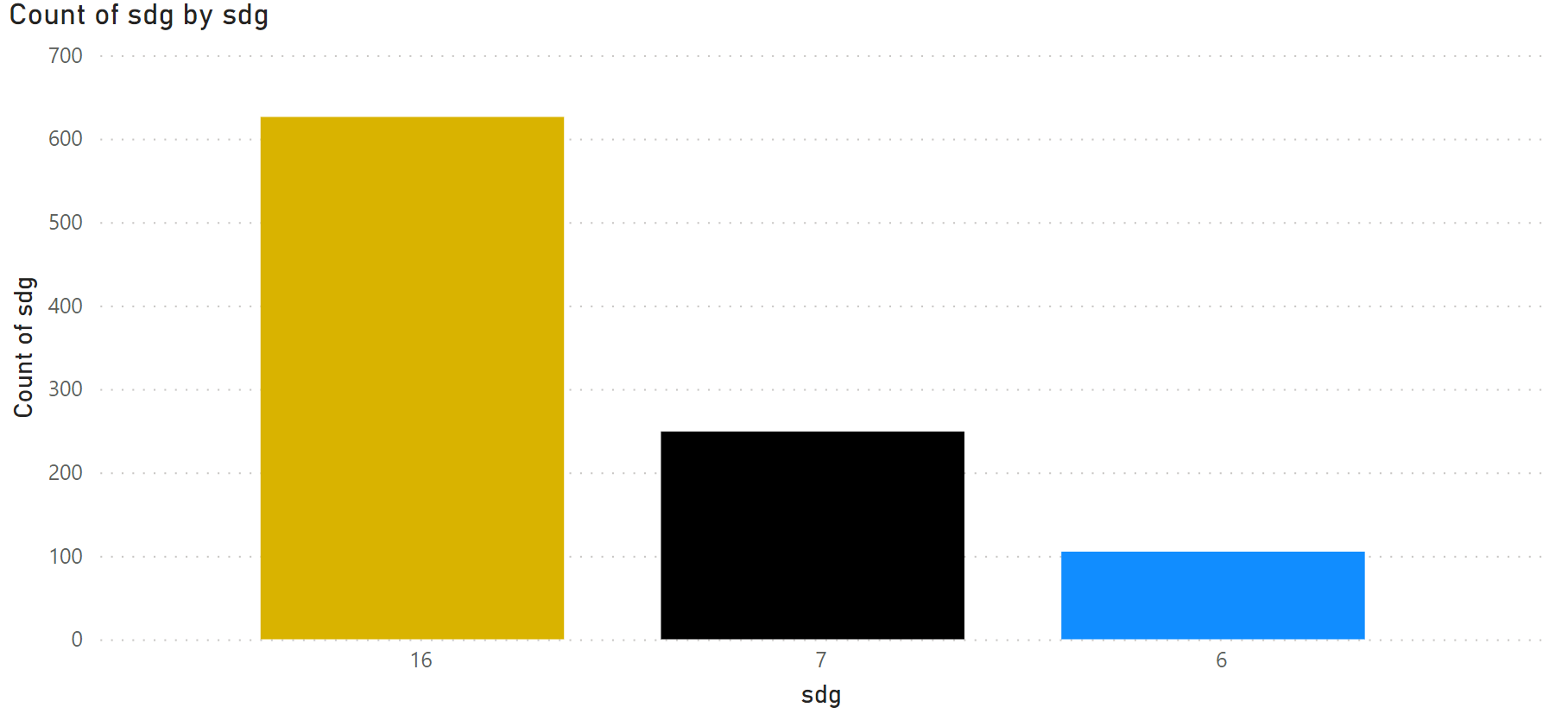
Description automatically generated with medium confidence

Resultado de la predicción con los valores de prueba

Dado que se utilizan ciertos registros de la tabla para entrenar el modelo, no tiene sentido hacer la predicción con esos datos dado que tendría 100% de precision, recall y f1-score.

Dicho esto, al comparar sus weighted avg y valores individuales de precisión, recall y f1-score, se puede afirmar aunque tienen buenos valores de predicción, los algoritmos de Random Forest y Logistic Regresion son mejores modelos para predecir si dado un texto, la categoría de ODS es una u otra.

Finalmente, al tener toda esta información se puede afirmar que, si al negocio no le afecta tener un menor valores variables individualmente en recall y precisión sino, tener un weighted avg bueno, se puede seleccionar Random Forest o Logistic Regresion. Sin embargo, si tiene alguna prioridad en predecir algún tipo de ODS especifico, puede que sea mas recomendable elegir Random Forest dado que generalmente, tiene precisión, recall y f1-score igual o mejores para cada ODS que Logistic Regresion.



# Parte 5: Mapa de actores

Para nuestro caso identificamos que una posible empresa que podría beneficiarse del modelo que creamos seria una empresa de energía y medio ambiente, para la cual se identifico el siguiente mapa de actores:

|  |  |  |  |
| --- | --- | --- | --- |
| Actor | Tipo de actor | Beneficio | Riesgo |
| Gobiernos locales y reguladores | Usuario-cliente | Apoya a mejorar los procesos de planeación participativa para el desarrollo a nivel territorial. | Si el modelo no tiene un buen desempeño, podemos estar creando e implementando políticas que no sean afines con el ODS |
| Comunidades locales | Beneficiado | Son escuchadas por los gobiernos locales en cuanto a sus necesidades para el desarrollo territorial | El problema puede ser tan especifico que no sea tomado en cuenta. |
| Dirección de ingeniería de sistemas y computación | Financiador | Mecanismo de toma de decisiones basado en información, menos sensible a la intuición del personal | En caso de que el modelo no funcione es dinero mal invertido y pudo dejarse de hacer un proyecto con mayor impacto y viabilidad. |

# Parte 6: Trabajo en equipo

|  |  |  |
| --- | --- | --- |
| Integrante | Rol | Tareas realizadas |
| Carlos Andrés García Gómez | Líder del proyecto  Líder de negocio | Repartición de actividades  Implementación del algoritmo de Random Forest |
| Juan Sebastián Pinzón Sampedro | Líder de datos | Implementación del algoritmo de Logistic Regression  Limpieza y preparación de los datos |
| Andres Felipe Garcia Bernal | Líder de analitica | Implementación del algoritmo de Decision Trees |