Universidad de los Andes
Ingeniería de Sistemas y Computación

Entrega 1 – Grupo 36

Integrantes:

Nombre: Sergio Oliveros **Código:** 202123159

Nombre: Camilo Daza **Código:** 201416461

Colab del Proyecto: https://colab.research.google.com/drive/1Wr3x-9NunlgCAGevXqzAaL_kVbDU0hEc?usp=sharing

Sección 1. (10%) Entendimiento del negocio y enfoque analítico.

Categoría	Descripción
Oportunidad/problema Negocio	El problema central es la dificultad y el consumo de recursos que conlleva analizar manualmente la gran cantidad de opiniones ciudadanas en lenguaje natural. La oportunidad radica en automatizar este proceso, relacionando las opiniones con los Objetivos de Desarrollo Sostenible (ODS) 3 (Salud y Bienestar), 4 (Educación de Calidad) y 5 (Igualdad de Género). Esto optimizaría el uso de los recursos, reduciría el tiempo de análisis y permitiría tomar decisiones más rápidas.
Objetivos y criterios de éxito desde el punto de vista del negocio:	• Desarrollar un modelo analítico capaz de procesar y categorizar opiniones ciudadanas de manera automática. • Relacionar correctamente las opiniones con los ODS 3, 4 y 5. • Reducir la necesidad de intervención manual en el análisis de opiniones.
Organización y rol dentro de ella que se beneficia con la oportunidad definida:	Las entidades públicas que colaboran con el UNFPA, responsables de formular políticas y evaluar las condiciones de salud, educación y equidad de género, se beneficiarían enormemente. A nivel organizacional, esto mejoraría la toma de decisiones informada, permitiendo asignar recursos de manera más eficiente para lograr los ODS.
Impacto que puede tener en Colombia este proyecto:	El impacto en Colombia sería significativo, ya que facilitaría la implementación de políticas más centradas en las necesidades reales de los ciudadanos, acelerando el progreso hacia los ODS.
Enfoque analítico. Descripción de la categoría de análisis (descriptivo, predictivo, etc.), tipo y tarea de aprendizaje e incluya las técnicas y algoritmos que propone utilizar.	**Categoría de análisis:** Predictivo, dado que se quiere predecir la clasificación de las opiniones de las personas con base en lo aprendido antes, logrando clasificar correctamente en uno de los 3 ODS. **Tipo de aprendizaje:** Aprendizaje supervisado, ya que se contaría con datos etiquetados para entrenar el modelo. **Tarea de aprendizaje:** Clasificación, específicamente de texto. **Técnicas/algoritmos:** Para resolver la tarea propuesta, se harán análisis por medio de 3 algoritmos distintos. - Random Forest - SVM - MLP

Sección 2. (20%) Entendimiento y preparación de los datos.

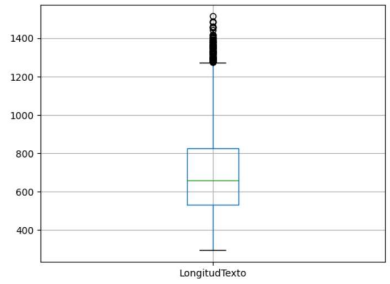
El análisis de los detalles realizados queda expuesto en el notebook adjunto del proyecto. A continuación, un resumen de los análisis realizados:

Perfilamiento y análisis de calidad de los datos

Variable “Textos_espanol”

Para iniciar se importaron los datos en un dataframe de pandas y se revisó de manera general su estructura, notando que se compone de dos columnas “Textos_espanol” y “sdg”, que corresponden al fragmento de texto que habla acerca de un objetivo de desarrollo sostenible y la segunda, que corresponde al ods que corresponde dicho fragmento de texto.

Empezando a entender la variable “textos espanol”, se quiso inspeccionar la longitud de los datos.

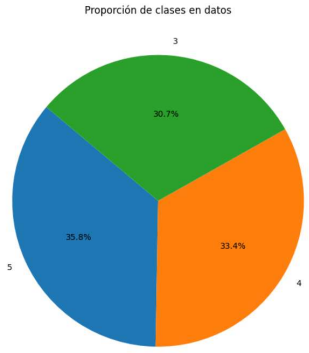


Del grafico de cajas y bigotes podemos ver que la mayoría de los textos tiene una longitud de entre 500 caracteres y 800. Sin embargo, se presentan unos cuantos outliers superan los rangos normales.

	LongitudTexto
count	4049.000000
mean	699.632502
std	228.988965
min	294.000000
25%	531.000000
50%	657.000000
75%	827.000000
max	1513.000000

Específicamente se encontraron 93 casos de textos por encima del rango normal. Sin embargo, dado que el enfoque principal en el procesamiento del lenguaje natural es la semántica de los textos, no su rango de valores, estos se dejaron como parte del dataset.

Variable “sdg”



Ahora, pasando a la segunda variable, “sdg”, se encuentra que este corresponde a un problema de clasificación balanceado, de forma que no es necesario emplear técnicas para contrarrestar este problema a la hora de desarrollar el modelo.

Análisis de calidad de los datos

- Análisis de Completitud

Al revisar las variables del dataframe vemos que ninguna tiene valores faltantes. Indicando que los datos están completos.

- Análisis de Unicidad

Ninguno de los textos se repite. De manera que cada uno de los datos son únicos.

- Análisis de Consistencia

Para confirmar la consistencia de los datos sería necesario revisar cada uno de los textos y confirmar que son semánticamente consistentes con el ODS que se encuentra en la variable “sdg”. Por cuestiones prácticas esto es inviable. Por lo que se decidió hacer fue tomar muestras al azar y confirmar que estas fueron consistentes. Los resultados indicaron que son consistentes.

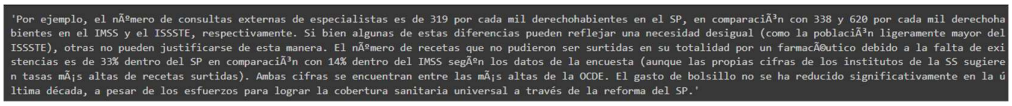
- Análisis de Validez

Para la variable *Textos_espanol* todas las entradas eran strings, confirmando que los textos eran en efecto textos. Sin embargo, estos textos contienen caracteres no pertenecientes al lenguaje español, de forma que es necesario corregir este problema. Por otro lado, al revisar la variable “sdg”:

Es posible notar que no hay clases que no pertenezcan a lo esperado. Indicando que la variable “sdg” es válida.

Transformación de los datos

Ajuste de texto incorrecto: Analizando los textos de la variable “Textos_espanol”, se pudo notar que algunas entradas de la variable contenían caracteres no propios del lenguaje español.

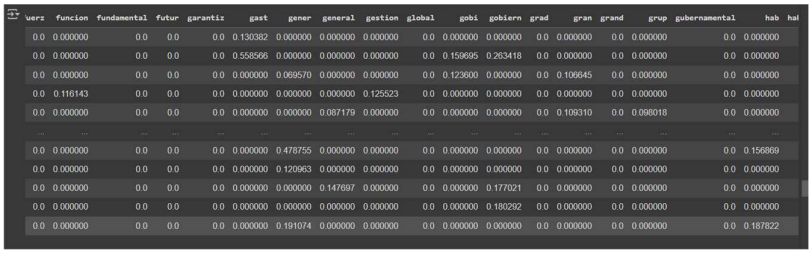


De forma que se aplicó un preprocesamiento para sustituir estos errores por los caracteres correspondientes.

Eliminación de caracteres y palabras vacías: Al realizar estos análisis, también se debe tener en cuenta que no todas las palabras contenidas aportan valor en el análisis, muchas de ellas son conectores o palabras comunes, al igual que números o signos de puntuación. Por lo anterior, se realizó la eliminación de todos estos. Para ello todos los caracteres fueron convertidos a minúsculas y, utilizando la librería *NLTK* se identificaron las palabras vacías del lenguaje español para poder eliminarlas.

Transformar palabras a su raíz en español: Por medio de NLTK se utilizó la clase *SnowballStemmer* debido a que su capacidad de encontrar la raíz de las palabras del lenguaje español, para así también reducir el número distinto de palabras, como, por ejemplo, la palabra corriendo a corr. Permitiendo al algoritmo apreciar características generales de los textos y evitando que se centre en palabras específicas con significados casi que idénticos.

Representación de textos: Una vez con los textos representados como listas de palabras raíz se realiza *Term Frequency – Inverse Document Frequency*. Esta es una técnica que permite determinar la importancia relativa de una palabra dentro de un texto perteneciente a un conjunto de textos. De manera que cada texto se puede representar como un vector en el cual cada entrada indica que tan importante es una palabra dada en dicho vector. Las entradas están entre 0 y 1, ayudando a la convergencia de los modelos. Cada vector tiene dimensión de 629. Las palabras seleccionadas para componer el vector se determinaron al determinar aquellas que tienen una frecuencia en los documentos mayor a 1.7%.



uerz	funcion	fundamental	futur	garantiz	gast	gener	general	gestion	global	gobi	gobiern	grad	gran	grand	grup	gubernamental	hab	hal
0.0	0.000000	0.0	0.0	0.0	0.130382	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.000000	
0.0	0.000000	0.0	0.0	0.0	0.558566	0.000000	0.000000	0.000000	0.0	0.159695	0.263418	0.0	0.000000	0.0	0.000000	0.0	0.000000	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.069570	0.000000	0.000000	0.0	0.123600	0.000000	0.0	0.106645	0.0	0.000000	0.0	0.000000	
0.0	0.116143	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.125523	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.000000	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.087179	0.000000	0.0	0.000000	0.000000	0.0	0.109310	0.0	0.098018	0.0	0.000000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.478755	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.156869	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.120963	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.000000	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.147697	0.000000	0.0	0.000000	0.177021	0.0	0.000000	0.0	0.000000	0.0	0.000000	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.180292	0.0	0.000000	0.0	0.000000	0.0	0.000000	
0.0	0.000000	0.0	0.0	0.0	0.000000	0.191074	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.187822	

Creación de Matriz: Se crea una matriz donde en las filas se tiene cada documento (en su forma de arreglo) y en las columnas la frecuencia relativa de cada palabra en los documentos.

Con esto se logró tratar los datos y transformarlos a la medida de los modelos de lenguaje natural utilizar.

Partición de datos: Finalmente se dividen los datos en entrenamiento y validación, con el fin de ser capaces de encontrar el modelo que mejor generalice los datos.

Sección 3. (20%) Modelado y evaluación.

El análisis detallado de los modelos se encuentra en el Collab adjunto.

Para resolver la tarea de aprendizaje automático propuesto, se utilizaron 3 algoritmos distintos para la construcción de los modelos y sus métricas asociadas, los algoritmos seleccionados fueron el Random Forest, el SVM y el NLP.

Random Forest (Camilo Daza)

Como primer modelo se usó el Random Forest el cual es un modelo de ensamble dado que combina múltiples árboles de decisión para realizar la predicción, permitiendo reducir el sobreajuste a los datos vs un árbol de decisión simple, y también el ser más estable antes pequeñas variaciones en los datos dado que combina múltiples árboles. De igual forma, este modelo presenta una buena robustez al ruido o valores atípicos dado que el resultado final es el resultado mayoritario de todos los árboles. Por último, este modelo muestra gran versatilidad al analizar tanto datos numéricos como categóricos sin realizar gran cantidad de transformaciones.

Para determinar los mejores hiperparámetros se realizó una búsqueda de grilla variando los siguientes hiperparámetros:

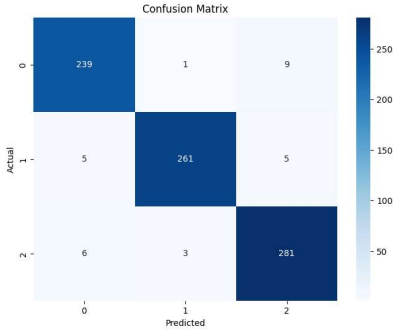
```
param_grid = {'n_estimators': [750, 1250], 'criterion': ["gini", "entropy", "log_loss"], "max_depth":[20, 50,None]}
```

Los hiperparámetros del mejor modelo fueron los siguientes:

```
{'criterion': 'entropy', 'max_depth': None, 'n_estimators': 1250}
```

	precision	recall	f1-score	support
3	0.96	0.96	0.96	249
4	0.98	0.96	0.97	271
5	0.95	0.97	0.96	290
accuracy			0.96	810
macro avg	0.96	0.96	0.96	810
weighted avg	0.96	0.96	0.96	810

Al evaluar un Random Forest sobre los datos de validación se obtuvieron los siguientes resultados:



Support Vector Machine (SVM) – (Sergio Oliveros)

Como segundo modelo, se usó el SVM, modelo que por definición busca el hiperplano que maximiza las distancia entre 2 categorías, lo cual es perfecto para los problemas de clasificación, siendo este un modelo que se encarga de alejar distintos tipos de información al máximo y clasificarlos. Cabe restaltar que la forma en que logra clasificar múltiples clases es mediante el uso de múltiples SVMs, haciéndolo un modelo de ensamble. La forma en que lo logra es mediante 2 métodso “One v One” o “One v Rest”; El “One v One” entrena modelos para diferenciar entre 2 clases específicas, mientras que el “One v Rest” se entrena modelos para diferenciar una clase del resto. Por último, este tipo de modelos suelen presentar mayor precisión cuando la data del problema está claramente separada, y es un buen complemento de los árboles, los cuales suelen ser más robustos en problemas con datos mixtos.

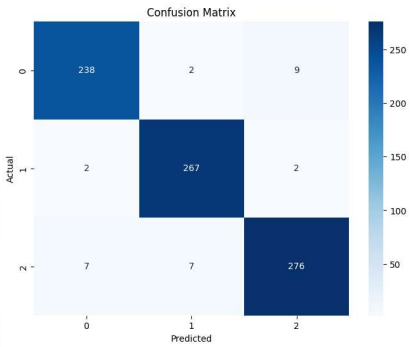
Para determinar los mejores hiperparámetros se realizó una búsqueda de grilla variando los siguientes hiperparámetros:

```
param_grid = {'C': [2, 3, 4], 'kernel': ["poly", "rbf", "sigmoid"], "gamma":["scale", "auto", 0.1, 10], "decision_function_shape" :["ovo", "ovr"],}
```

Los hiperparámetros del mejor modelo fueron los siguientes:

```
{'C': 3, 'decision_function_shape': 'ovo', 'gamma': 'scale', 'kernel': 'rbf'}
```

	precision	recall	f1-score	support
3	0.96	0.96	0.96	249
4	0.97	0.99	0.98	271
5	0.96	0.95	0.96	290
accuracy			0.96	810
macro avg	0.96	0.96	0.96	810
weighted avg	0.96	0.96	0.96	810



Multi-Layer Perceptron (MLP) – (Sergio Oliveros)

El tercer modelo escogido fue el MLP, este modelo es la forma más básica de una red neuronal, la cual es capaz de aprender relaciones no lineales en los datos. Este tipo de modelo suele tener mayor precisión y rendimiento cuando hay suficiente data (suele necesitar más información para entrenarse) y también complementa muy bien a los dos modelos anteriormente usados, dado que puede manejar datos no lineales de manera más directa. De igual forma, este modelo al tener múltiples capas ocultas logra capturar relaciones que otros modelos no podrían, a la vez que es un tipo de modelo que se ajusta a una gran cantidad de problemas.

Para determinar los mejores hiperparámetros se realizó una búsqueda de grilla variando los siguientes hiperparámetros:

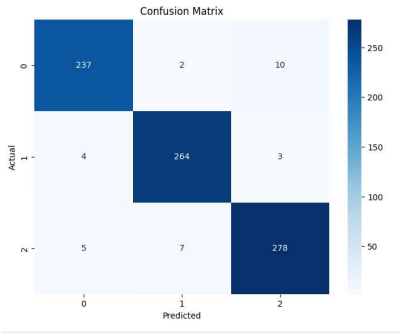
```
param_grid = {'hidden_layer_sizes': [(512, 256, 128, 64, 32), (512, 256, 128, 64, 32, 16), (512, 128, 32, 16), (256, 64, 32, 16), (512, 128, 32), (256, 64, 16)], 'activation': ["logistic", "tanh", "relu"], "learning_rate":["constant", "adaptive"], }
```

Los hiperparámetros del mejor modelo fueron los siguientes:

```
{'activation': 'relu', 'hidden_layer_sizes': (512, 256, 128, 64, 32, 16), 'learning_rate': 'constant'}
```

Al evaluar un MLP sobre los datos de validación se obtuvieron los siguientes resultados:

	precision	recall	f1-score	support
3	0.96	0.95	0.96	249
4	0.97	0.97	0.97	271
5	0.96	0.96	0.96	290
accuracy			0.96	810
macro avg	0.96	0.96	0.96	810
weighted avg	0.96	0.96	0.96	810



Los tres modelos de estudio presentaron muy buenos resultados en las métricas de precisión, recall, f-score y accuracy.

Sección 4. (20%) Resultados.

a. Descripción de los resultados obtenidos

Después de analizar los resultados obtenidos por los modelos en la etapa de validación, es posible afirmar que se cuenta con 3 modelos capaces de resolver el problema de clasificación de textos un Objetivos de Desarrollo Sostenible.

El primer modelo, Random Forest, logró obtener una precisión promedio de 96%, una cobertura promedio de 96% y una f-medida de 96%. Todas métricas considerablemente buenas.

De manera similar los modelos SVM y MLP también lograron obtener una precisión promedio de 96%, una cobertura promedio de 96% y una f-medida de 96%. Por lo que se podría considerar que todos los modelos son igual de efectivos para la tarea. Según lo anterior, es posible afirmar que se cuenta con un modelo analítico capaz de procesar y categorizar las opiniones ciudadanas de forma automática y poder relacionarlas directamente con alguno de los ODS 3,4 y 5. Adicionalmente, esto reducirá en gran medida la intervención manual en el análisis de opiniones y permitirá tomar decisiones con mayor velocidad.

b. Incluir el análisis de las palabras identificadas

Para hacer un análisis de las palabras más importantes lo que se hizo fue tomar los documentos que fueron clasificadas en cada ODS del conjunto de validación. Después de ello para cada ODS se sumaron los vectores que representaban los documentos. Para finalmente, identificar las posiciones de los vectores (palabras) con los valores más altos. El resultado fue el siguiente:

```
Palabras más importantes ods3: ['salud' 'atencion' 'servici' 'medic' 'mental' 'sanitari' 'pacient'
 'sistem' 'calid' 'enfermedad']
Palabras más importantes ods4: ['educ' 'estudi' 'escuel' 'docent' 'aprendizaj' 'evalu' 'program' 'alumn'
 'enseñ' 'sistem']
Palabras más importantes ods5: ['mujer' 'gener' 'derech' 'hombr' 'trabaj' 'iguald' 'pued' 'polit' 'pais'
 'niñ']
```

Esto lo que nos indica son las palabras más importantes para clasificar los documentos, es decir si un documento contiene estas palabras muy probablemente será clasificado en el ODS correspondiente. Adicionalmente, cabe notar que estos tienen sentido, el ODS 3 trata de salud y bienestar, y las palabras que prioriza el clasificador para este ODS tienen que ver con salud. Lo mismo para el ODS 4 que trata de educación de calidad y las palabras relacionadas tratan de educación. Finalmente, el ODS 5 es sobre igualdad de género y, como se puede ver, los términos son relacionados al género.

Ahora bien, estas palabras y los clasificadores desarrollados son de gran importancia para el Fondo de Poblaciones de las Naciones Unidas pues les permite identificar rápida el objetivo de desarrollo sostenible al cual se relaciona un texto de manera que es posible dividir los documentos y así agilizar el proceso de recopilación de opiniones por parte de la ciudadanía. Aumentando la eficiencia de la organización.

c. Entregar los datos de prueba compartidos

El modelo seleccionado para generar las predicciones sobre el conjunto de prueba fue Random Forest. La razón por la que se descartó MLP es debido a que, al sumar los valores por fuera de la diagonal en las matrices de confusión, MLP obtenía en total 31 datos clasificados incorrectamente. Mientras que SVM y Random Forest alcanzaban valores de 29.

Por otro lado, para elegir entre Random Forest y SVM, lo que se hizo fue generar las predicciones de ambos modelos para el conjunto de prueba y luego se revisó en qué casos las predicciones diferían, a partir de ello se tomó el modelo que parecía hacer mejores predicciones. Los resultados quedaron almacenados en el archivo *rfTestPredictions.xlsx*.

d. Generar un video de máximo 5 minutos. Subido al padlet del proyecto.

Sección 5. (10%)

En el caso de estudio, la principal organización en busca de utilizar de la mejor manera es el **Fondo de Poblaciones de las Naciones Unidas (UNFPA)**, en donde a continuación se presenta el mapa de Actores con sus beneficios y riesgos asociados:

Rol	Tipo de actor	Beneficio	Riesgo
Ciudadanos	Beneficiados del fondo	Permite que sus opiniones sean tomadas en cuenta en las decisiones de los gobiernos a nivel internacional, mejorando la implementación de políticas públicas asociadas a los ODS.	Si el modelo tiene un mal desempeño clasificando sus opiniones, las problemáticas principales de los ciudadanos podrían no ser atendidas. Lo cual puede llevar también a desconfianza de los ciudadanos al completar esta información.
Gobiernos	Cliente	Obtienen información precisa sobre los problemas de sus ciudadanos, lo que facilita la implementación de políticas alineadas con los ODS.	Si la información no es precisa, podrían implementarse políticas ineficaces, desviando recursos y tiempo en proyectos que no reflejan las necesidades ciudadanas.
Dirección de Temas sociales	Cliente	Permite identificar correctamente las necesidades de las personas	Si el modelo genera resultados inexactos, se podrían priorizar proyectos de menor

		en los 3 ODS, facilitando la priorización de proyectos	importancia, generando un malgasto de los recursos.
Equipo de Analítica de datos	Proveedor	Asegura el cumplimiento de la creación de productos tecnológicos avanzados que permiten procesar un gran volumen de datos, a la vez que entregar información útil para la toma de decisiones de distintos actores	Sino se maneja adecuadamente la privacidad de la información obtenida, se puede llegar a generar violación sobre los derechos de los ciudadanos y caída en la reputación de la organización
Organizaciones bilaterales	Financiador	Se facilita el conceder financiación para las iniciativas alineadas con sus objetivos específicos.	Si el modelo muestra problemas de clasificación, se podrían estar financiado iniciativas no prioritarias para los ciudadanos, generando desconfianza hacia las decisiones tomadas por las bilaterales a financiar.

Sección 6. (8%) Trabajo en equipo

Roles del proyecto:

Líder del proyecto y Negocio: Camilo Daza

Líder de datos y analítica: Sergio Oliveros

Reuniones

1. **Planeación:** Lunes 2 de Septiembre

El equipo se reunió para llevar a cabo la primera sesión de planeación del proyecto. En esta sesión mayoritariamente nos dedicamos a entender el problema planteado, hicimos una lluvia de ideas sobre las posibles soluciones y establecimos los roles del equipo. Se decidió que Camilo se encargaría principalmente de asignar las actividades, enfocarse en las tareas de entendimiento del problema y asegurar el cumplimiento de los objetivos del problema planteado. Al igual que se decidió que Sergio liderará el análisis de datos, dejando el collab listo y gestionará las actividades de análisis y elección del modelo.

2. **Seguimiento:** Jueves 5 de Septiembre

El equipo se reunió para solucionar dudas del análisis realizado hasta el momento y discutir sobre los modelos a seleccionar. El líder de datos propuso distintos tipos de modelo y se escogieron para ser analizados, al igual se establecieron los indicadores a tener en cuenta para en un futuro escoger el mejor modelo de los 3 (En un futuro el líder de analítica tuvo que escoger con análisis adicionales)

3. **Finalización:** Sábado 7 de Septiembre

El equipo se reunió el sábado para entender resultados finales obtenidos y sus implicaciones, coordinar el cierre del documento, finalizar la presentación y realizar el video final de esta entrega del proyecto.

Carga real vs carga final

Al inicio del proyecto en las sesiones de planeación se planteó una distribución equitativa del 50%-50%, pero con actividades asociadas a su rol. En este sentido, Sergio se encargó más del análisis de la información y Camilo del entendimiento del problema. Lo anterior, conllevó que, al momento de realizar las actividades asociadas, la parte analítica, de entendimiento de los datos, perfilamiento e ideación de modelos gastara más tiempo que el de otras actividades asociadas al entendimiento del problema y a la resolución de este. Estimamos que al final del proyecto, Sergio realizó un 60% de las actividades que requirieron mayor tiempo, esto teniendo en cuenta que el realizó 2/3 modelos y una parte importante del perfilamiento de la data, lo cual dejó con actividades de menor peso a Camilo. Estimamos que Sergio trabajó alrededor de 6horas en el proyecto y Camilo 4h. Por lo anterior, si dividiéramos los 100 puntos del proyecto, Sergio tendría 60 puntos y Camilo 40 puntos. Se espera que para próximas iteraciones se ajusten las cargas realizadas del proyecto.