

# PROYECTO ETAPA 1

## CONSTRUCCIÓN DE MODELOS DE ANALITICA DE TEXTOS

Elaborado por:

Andres Felipe Guerrero Sarmiento – 202015143

Julian Camilo Rivera — 202013338

Ricardo Andrés Sanchez Alvarez — 202014809

Institución:

Universidad De Los Andes

Curso:

Inteligencia de Negocios

Profesor:

Haydemar Núñez

## **TABLA DE CONTENIDO**

- Descripción
- Entendimiento del negocio y enfoque analítico
- Entendimiento y preparación de los datos
- Modelado y evaluación
- Resultados
- Mapa de actores relacionado con el producto de datos creado
- Trabajo en equipo
- Encuentros

### **Descripción**

En esta etapa desempeñan principalmente el rol de científico de datos. Cada grupo debe trabajar con reseñas de sitios turísticos y calificaciones dadas a los mismos para construir un modelo analítico que permita realizar la calificación automática de nuevas reseñas, con un alto nivel de precisión y de sensibilidad (recall). De igual manera, deben describir las palabras seleccionadas para representar las reseñas, las cuales corresponden a las variables a utilizar en el modelo analítico.

## Entendimiento del negocio y enfoque analítico

*a. Definición de los objetivos y criterios de éxito desde el punto de vista del negocio.*

### **Objetivos:**

- Automatizar el proceso de calificación de reseñas de sitios turísticos.
- Aplicar la metodología de analítica de textos para la construcción de soluciones de analítica alineadas con los objetivos de organizaciones en un contexto de aplicación.
- Planear la interacción con un grupo interdisciplinario para identificar usuarios y posibles herramientas a desarrollar que faciliten la interacción del resultado del modelo analítico desarrollado.

### **Criterios de éxito:**

- Alta precisión y sensibilidad en la calificación automática de nuevas reseñas.
- Validación y mejora continua del modelo analítico y la aplicación desarrollada.
- Desarrollo de una aplicación útil y eficiente para la organización y sus stakeholders.

El hecho de realizar este proyecto de forma satisfactoria puede generar un impacto significativo en el sector del turismo colombiano, principalmente en todos aquellos negocios como hoteles y restaurantes de estas zonas, con altos números de visitantes turistas, podrán identificar áreas de mejora y desarrollar estrategias para aumentar la popularidad de los destinos turísticos. con esto se benefician tanto los negocios, como la experiencia del turista.

*b. Descripción del enfoque analítico para alcanzar los objetivos del negocio.*

El método analítico propuesto emplea avanzadas herramientas de procesamiento de lenguaje natural y aprendizaje automático para crear un sistema que simplifique la evaluación de opiniones sobre destinos turísticos. Con esta estrategia, las empresas del sector turismo podrán analizar de manera eficaz y precisa las reseñas de sus clientes, detectar patrones y áreas de mejora, y tomar decisiones informadas para mejorar la experiencia del turista

Al utilizar algoritmos de análisis de sentimientos y modelos de clasificación, se logrará categorizar automáticamente las reseñas según su tono emocional y prever la calificación asociada, lo que facilitará la identificación de aspectos específicos que influyen positiva o negativamente en la percepción.

*c. Nombres de los estudiantes del curso de estadística con los cuales va a interactuar para validar el enfoque que le está dando el proyecto*

Barak Valderrama y Laura Velandia son los estudiantes del curso de estadística con los que el grupo ha interactuado para validar los enfoques del proyecto en cuestión.

Con ellos se ha tenido entre una y dos reuniones por semana, se tiene contacto con ellos a través de WhatsApp y las reuniones se hacen por medio de Zoom.

Oportunidad/problema Negocio	Automatización del proceso de calificación de reseñas de sitios turísticos para mejorar la toma de decisiones en el sector turismo.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Análisis de Sentimientos: Utilizar técnicas de procesamiento de lenguaje natural (NLP) para analizar el sentimiento de las reseñas de los turistas hacia los sitios turísticos. Esto incluiría la tokenización de texto, eliminación de stop words y normalización. para elegir los mejores algoritmos quisimos implementar los 5 propuestos: modelo Support Vector Machine, modelo AdaBoostClassifier, modelo Naive Bayes y modelo KNN.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia (COTELCO), así como cadenas hoteleras como Hilton, Hoteles Estelar, Holiday Inn y hoteles pequeños ubicados en diferentes municipios de Colombia se beneficiarían directamente al mejorar la calidad de las decisiones relacionadas con el turismo.
Contacto con experto externo al proyecto y detalles de la planeación	Se puede establecer contacto con expertos en análisis de sentimientos, procesamiento de lenguaje natural y aprendizaje automático para obtener asesoramiento sobre las mejores prácticas y enfoques para abordar el problema. Se decide reunirse con los compañeros de estadística entre 1 y 2 veces por semana por todo el tiempo del desarrollo del proyecto, para notificar avances y evolución del desarrollo del proyecto.

## Entendimiento y preparación de los datos

El primer paso realizado antes de comenzar a desarrollar el proyecto fue una lectura y análisis de los datos y tipos de datos sobre los que íbamos a trabajar a lo largo de este como se puede ver a continuación:

data_t		
	Review	Class
0	Nos alojamos en una casa alquilada en la ciuda...	4
1	La comida está bien, pero nada especial. Yo te...	3
2	En mi opinión, no es una como muchos usuarios ...	3
3	esta curiosa forma que asemeja una silla de mo...	4
4	Lo mejor era la limonada. Me gusto la comida d...	2
...	...	...
7870	El motivo de mi estancia fue porque vine a un ...	3
7871	Es difícil revisar el castillo porque apenas p...	3
7872	Si vas a Mérida no puedes perderte de este lug...	5
7873	Este imperdible sitio, que lleva el nombre del...	5
7874	Festejando Día del Amor y Amistad\n\nTe remont...	3

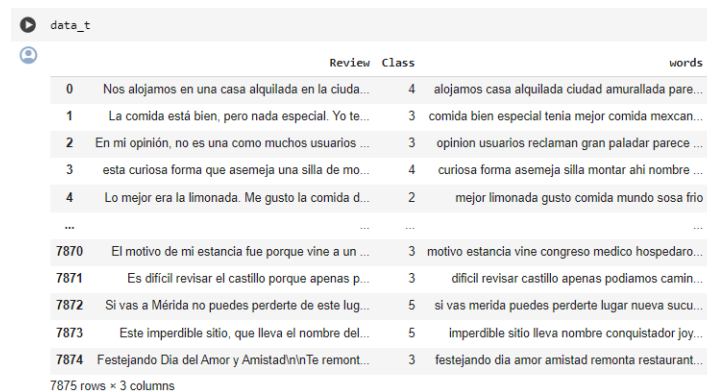
7875 rows x 2 columns

Con esto y apoyándonos en el enunciado brindado logramos comprender que trabajaríamos con datos que consisten en reseñas y sus respectivas calificaciones de diferentes cadenas hoteleras y hoteles distribuidos por el país.

Teniendo esto en cuenta nuestro siguiente paso fue la preparación de los datos para lo que implementamos las diferentes prácticas vistas a lo largo del curso como lo son la limpieza de los datos, la tokenización y la normalización. Lo anterior mencionado podemos apreciarlo de una mejor forma a continuación:

- Limpieza de los datos

Este proceso se realizó con el objetivo de dejar el archivo en texto plano, eliminar caracteres especiales y pasar todo a minúscula, cabe aclarar que de igual manera se realizó un proceso de corrección de las contracciones presentes en los textos como prerequisite del siguiente paso. A continuación podemos observar los resultados obtenidos de este procesamiento (Mayo, 2020):

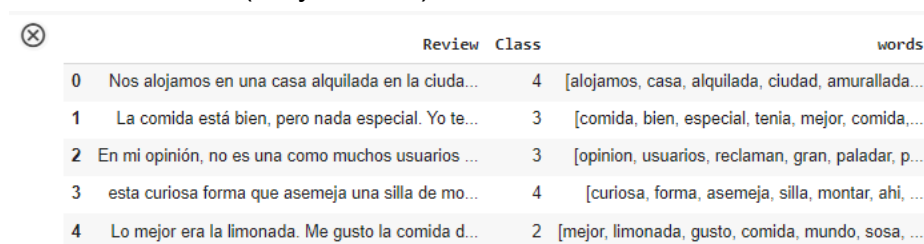


	Review	Class	words
0	Nos alojamos en una casa alquilada en la ciuda...	4	alojamos casa alquilada ciudad amurallada pare...
1	La comida está bien, pero nada especial. Yo te...	3	comida bien especial tenia mejor comida mexcan...
2	En mi opinión, no es una como muchos usuarios ...	3	opinion usuarios reclaman gran paladar parece ...
3	esta curiosa forma que asemeja una silla de mo...	4	curiosa forma asemeja silla montar ahi nombre ...
4	Lo mejor era la limonada. Me gusto la comida d...	2	mejor limonada gusto comida mundo sosa frio
...	...	...	...
7870	El motivo de mi estancia fue porque vine a un ...	3	motivo estancia vine congreso medico hospedaro...
7871	Es difícil revisar el castillo porque apenas p...	3	dificil revisar castillo apenas podiamos camin...
7872	Si vas a Mérida no puedes perderte de este lug...	5	si vas merida puedes perderte lugar nueva sucu...
7873	Este imperdible sitio, que lleva el nombre del...	5	imperdible sitio lleva nombre conquistador joy...
7874	Festejando Dia del Amor y Amistad/n/nTe remont...	3	festejando dia amor amistad remonta restaurant...

7875 rows x 3 columns

- Tokenización

El objetivo de este proceso fue el dividir las frases u oraciones en palabras con el fin de desglosar las palabras correctamente para su posterior análisis, esto se puede evidenciar a continuación (Mayo, 2020):



	Review	Class	words
0	Nos alojamos en una casa alquilada en la ciuda...	4	[alojamos, casa, alquilada, ciudad, amurallada...
1	La comida está bien, pero nada especial. Yo te...	3	[comida, bien, especial, tenia, mejor, comida,...
2	En mi opinión, no es una como muchos usuarios ...	3	[opinion, usuarios, reclaman, gran, paladar, p...
3	esta curiosa forma que asemeja una silla de mo...	4	[curiosa, forma, asemeja, silla, montar, ahi, ...
4	Lo mejor era la limonada. Me gusto la comida d...	2	[mejor, limonada, gusto, comida, mundo, sosa, ...

- Normalización

En este paso se realizó la eliminación de prefijos y sufijos junto a un proceso de lematización el cual consiste en dada una forma flexionada de una palabra se halla el lema correspondiente a esta como se observa a continuación:

data\_t

	Review	Class	words
0	Nos alojamos en una casa alquilada en la ciuda...	4	[aloj, cas, alquil, ciud, amurall, pareci, tan...
1	La comida está bien, pero nada especial. Yo te...	3	[com, bien, especial, teni, mejor, com, mexc, ...
2	En mi opinión, no es una como muchos usuarios ...	3	[opinion, usuari, reclam, gran, palad, parec, ...
3	esta curiosa forma que asemeja una silla de mo...	4	[curios, form, asemej, sill, mont, ahi, nombr,...
4	Lo mejor era la limonada. Me gusto la comida d...	2	[mejor, limon, gust, com, mund, sos, fri]
...	...	...	...
7870	El motivo de mi estancia fue porque vine a un ...	3	[motiv, estanci, vin, congr, medic, hosped, lu...
7871	Es difícil revisar el castillo porque apenas p...	3	[difícil, revis, castill, apen, podi, camin, s...
7872	Si vas a Mérida no puedes perderte de este lug...	5	[si, ir, mer, pued, perdert, lug, nuev, sucurs...
7873	Este imperdible sitio, que lleva el nombre del...	5	[imperd, siti, llev, nombr, conquest, joy, urb...
7874	Festejando Día del Amor y Amistad\n\nTe remont...	3	[festej, dia, amor, amist, remont, restaur, ca...

7875 rows × 3 columns

Como último paso realizado, se separó la variable predictora y los textos que se van a utilizar, quedando de la siguiente manera:

```
data_t['words_united'] = data_t['words'].apply(lambda x: ' '.join(map(str, x)))
data_t
```

	Review	Class	words	words_united
0	Nos alojamos en una casa alquilada en la ciuda...	4	[aloj, cas, alquil, ciud, amurall, pareci, tan...	aloj cas alquil ciud amurall pareci tanto segu...
1	La comida está bien, pero nada especial. Yo te...	3	[com, bien, especial, teni, mejor, com, mexc, ...	com bien especial teni mejor com mexc unid mar...
2	En mi opinión, no es una como muchos usuarios ...	3	[opinion, usuari, reclam, gran, palad, parec, ...	opinion usuari reclam gran palad parec ser pa...
3	esta curiosa forma que asemeja una silla de mo...	4	[curio, form, asemej, sill, mont, ahi, nombr, ...	curio form asemej sill mont ahi nombr icon ciu...
4	Lo mejor era la limonada. Me gusto la comida d...	2	[mejor, limon, gust, com, mund, sos, fri]	mejor limon gust com mund sos fri
...	...	...	...	...
7870	El motivo de mi estancia fue porque vine a un ...	3	[motiv, estanci, vin, congr, medic, hosped, lu...	motiv estanci vin congr medic hosped lug insta...
7871	Es difícil revisar el castillo porque apenas p...	3	[difícil, revis, castill, apir, podi, camin, s...	difícil revis castill apir podi camin sofoc ca...
7872	Si vas a Mérida no puedes perderte de este lug...	5	[si, ir, mer, pued, perdert, lug, nuev, sucurs...	si ir mer pued perdert lug nuev sucursal mas a...
7873	Este imperdible sitio, que lleva el nombre del...	5	[imperd, siti, llev, nombr, conquest, joy, urb...	imperd siti llev nombr conquest joy urbanasu a...
7874	Festejando Día del Amor y Amistad\n\nTe remont...	3	[festej, dia, amor, amist, remont, restaur, ca...	festej dia amor amist remont restaur cafeteri ...

7875 rows × 4 columns

## Modelado y evaluación

Tal como se presentó en la sección anterior, los algoritmos desarrollados para el proyecto fueron Random Forest, Support Vector Machine, AdaBoostClasifier, Naive Bayes y KNN para los cuales a continuación analizaremos los resultados obtenidos. Antes de comenzar el análisis dejaremos en claro que para cada uno de estos modelos tendremos en cuenta las métricas de precisión, recall, f1-score, support, accuracy, macro avg y weighted avg:

- **Random Forest**

Para este modelo logramos obtener los siguientes resultados:

Reporte de clasificación en datos de prueba:				
	precision	recall	f1-score	support
1	0.61	0.28	0.39	163
2	0.45	0.45	0.45	226
3	0.43	0.28	0.34	319
4	0.39	0.43	0.41	389
5	0.55	0.73	0.63	478
accuracy			0.48	1575
macro avg	0.49	0.43	0.44	1575
weighted avg	0.48	0.48	0.46	1575

Con esta implementación logramos obtener un valor de accuracy del 48% lo que nos indica que este modelo logra clasificar correctamente el 48% de las instancias en los datos de prueba. Además, en cuanto a las diferentes métricas evaluadas la precisión más alta se logró para la clase 1 con un valor de 0.61 lo que nos indica que el 61% de las instancias clasificadas como clase 1 realmente son clase 1, el recall más alto se logró para la clase 5 con un valor de 0.73 lo que nos indica que el 73% de todas las instancias de la clase 5 en los datos de prueba fueron identificados correctamente, el f1-score más alto se encuentra en la clase 5 de igual manera con un valor de 0.63 lo que nos indica un buen equilibrio entre precisión y recall para esta clase.

Con lo anteriormente mencionado llegamos a la conclusión que este modelo parece tener dificultades para clasificar correctamente algunas clases, especialmente las que hacen parte de las minoritarias como la clase 1 y 3.

- **Support Vector Machine**

Por otro lado, para este modelo obtuvimos los siguientes resultados:

Reporte de clasificación en datos de prueba:				
	precision	recall	f1-score	support
1	0.57	0.40	0.47	163
2	0.40	0.40	0.40	226
3	0.44	0.38	0.41	319
4	0.42	0.47	0.44	389
5	0.62	0.68	0.65	478
accuracy			0.50	1575
macro avg	0.49	0.47	0.47	1575
weighted avg	0.50	0.50	0.49	1575

En el caso de este modelo logramos aumentar el valor de accuracy del modelo al 50%, la precisión más alta fue la clase 5 con 62%, el recall más alto fue de 68% y el f1-score más alto fue de 0.65, con esto logramos observar un aumento positivo en los valores de las diferentes métricas lo que nos lleva a concluir que basándonos en los resultados de los 2 modelos que hemos visto hasta el momento, el modelo Support Vector Machine es el más apto.

- **AdaBoostClassifier**

Para el caso de este modelo, los resultados fueron:

Reporte de clasificación en datos de prueba:				
	precision	recall	f1-score	support
1	0.78	0.04	0.08	163
2	0.36	0.28	0.32	226
3	0.27	0.14	0.19	319
4	0.36	0.40	0.38	389
5	0.46	0.76	0.57	478
accuracy			0.40	1575
macro avg	0.45	0.33	0.31	1575
weighted avg	0.42	0.40	0.36	1575

Con este modelo obtuvimos un accuracy de 40%, el valor más alto de precision fue del 78% para la clase 1, el recall mayor fue de 76% para la clase 5 y el f1-score más alto fue de 0.57 para la clase 5.

- **Naive Bayes**

Reporte de clasificación en datos de prueba para Bernoulli Naive Bayes:				
	precision	recall	f1-score	support
1	0.55	0.41	0.47	163
2	0.41	0.38	0.40	226
3	0.36	0.26	0.30	319
4	0.39	0.32	0.35	389
5	0.52	0.75	0.61	478
accuracy			0.46	1575
macro avg	0.44	0.42	0.43	1575
weighted avg	0.44	0.46	0.44	1575

En este caso se logró un valor de accuracy del 46%, la precisión más alta fue de 55% para la clase 1, el recall mayor fue de 75% para la clase 5 y el f1-scores mayor fue de 0.61 para la clase 5.

- **KNN**

Reporte de clasificación en datos de prueba para KNN:				
	precision	recall	f1-score	support
1	0.49	0.30	0.37	163
2	0.37	0.28	0.32	226
3	0.36	0.25	0.30	319
4	0.35	0.40	0.37	389
5	0.48	0.64	0.55	478
accuracy			0.41	1575
macro avg	0.41	0.37	0.38	1575
weighted avg	0.41	0.41	0.40	1575

Finalmente para el último modelo se obtuvo un valor de accuracy del 41%, la precisión mayor fue de 49% para la clase 1, el recall más alto fue de 64% para la clase 5 y el f1-score mayor fue de 0.55.



## Resultados

Ahora que ya hemos observado las diferentes métricas obtenidas para los modelos hablaremos sobre la importancia de estas para la organización.

Los resultados anteriormente mencionados son muy importantes debido a que por ejemplo un alto valor de precisión garantiza que las recomendaciones de sitios turísticos proporcionadas a los turistas sean confiables y precisas, lo que mejora la satisfacción del cliente y fomenta la lealtad hacia la marca. Por otro lado, un alto valor de recall asegura que el modelo pueda identificar la mayor cantidad posible de reseñas relevantes, lo que permite a la organización abordar de manera proactiva las áreas de mejora y maximizar la experiencia del cliente.

Finalmente, podemos afirmar que el modelo de Support Vector Machine fue el que obtuvo un mejor desempeño por lo que sería el más adecuado para usar en este caso para realizar las clasificaciones debido al rendimiento anteriormente mostrado lo que conlleva a una buena capacidad de clasificar automáticamente las reseñas de sitios turísticos con alta precisión y sensibilidad, permitiendo a la organización identificar áreas de mejora y desarrollar estrategias efectivas para aumentar la popularidad y la calificación de los sitios turísticos, lo que en última instancia contribuye al crecimiento del turismo en Colombia.

## Mapa de actores relacionado con el producto de datos creado

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
turista	Beneficiado	Experiencias turísticas más personalizadas y agradables gracias a la mejora en la oferta de servicios y destinos turísticos basada en la información recopilada y analizadas.	Posible pérdida de privacidad o manipulación de datos personales si no se garantiza una adecuada protección y manejo de la información
Dueño de un hotel afiliado a COTELCO	Beneficiado	Beneficio: Acceso a información valiosa sobre las preferencias y opiniones de los clientes, lo que permite personalizar la experiencia del huésped	Riesgo: Sobrecarga de trabajo o desconfianza en el sistema automatizado, lo que puede afectar el normal desarrollo de las actividades operativas
Director de Turismo en el Ministerio de Comercio, Industria y Turismo de Colombia	Cliente	Beneficio: Obtención de información detallada sobre las preferencias y opiniones de los turistas hacia los destinos turísticos colombianos	Riesgo: Dependencia excesiva del modelo analítico para la toma de decisiones, lo que podría llevar a descuidar otras fuentes de información relevantes, así como posibles desafíos éticos en el uso de datos sensibles o decisiones.
Analista de Datos en el Ministerio de Comercio, Industria y Turismo de Colombia	Proveedor	Desarrollo y mantenimiento del modelo analítico, lo que permite generar información significativa para mejorar la competitividad del sector turístico colombiano.	Falta de disponibilidad o calidad de los datos, lo que puede afectar la precisión y eficacia del modelo analítico. Generando información imprecisa y falsa.
Departamento de Finanzas del Ministerio de Comercio, Industria y Turismo de Colombia	Financiador	Mejora de la rentabilidad y eficiencia operativa a través de estrategias proporcionadas por el modelo analítico, lo que aumenta el retorno de la inversión en tecnología y en datos.	Riesgo: Inversión inicial significativa en la implementación del modelo y la infraestructura necesaria, con la posibilidad de no obtener los resultados esperados en el corto plazo.

## Trabajo en equipo

Estudiante	Rol	Tareas	Tiempo	Algoritmo	Retos	Soluciones
Julián Rivera	Líder de proyecto, Líder de negocio	Entender las necesidades del negocio. Hacer acuerdos con el grupo de estudiantes de estadística. Definir las actividades a realizar.	10 horas	SVM, Random Forest, KNN	Entender las necesidades reales del negocio. Atomizar las actividades a realizar.	Realizar una lectura detallada del caso de negocio, y dividir en pasos las entregas necesarias.
Andrés Guerrero	Líder de datos	Estudio, entendimiento y preparación de los datos.	6 horas	AdaBoostClassifier	Hacer el estudio de la calidad de los datos. Buscar, entender e implementar las tecnologías necesarias para la limpieza y preparación de los datos.	Utilizar los recursos brindados en la clase, IA e información en línea para configurar las herramientas y hacer la verificación de la correcta transformación y preparación de los datos.
Ricardo Sánchez	Líder de analítica	Estudio, entendimiento y análisis de los resultados obtenidos.	6 horas	Naive Bayes	Hacer el estudio de los resultados de los modelos para comprobar la calidad de los mismos y los impactos de estos resultados sobre los objetivos del negocio.	Utilizar los objetivos de negocio y los resultados de calidad de los algoritmos para poder hacer una relación más clara entre estos.

## Encuentros

Reunión	
Fecha	Descripción del encuentro
4/04/24	Reunión inicial de lanzamiento y socialización de las actividades a realizar por los estudiantes de business intelligence a los estudiantes que están viendo estadística.
11/04/24	Reunión para exponer los resultados obtenidos y socializar las actividades asociadas a la realización del aplicativo web. Primera presentación de los mocks del aplicativo.
18/04/24	Reunión para exponer el resultado final del aplicativo web, pulir detalles y estética del mismo.

## Referencias

Mayo, M. (2020, March 12). Preprocesamiento de datos de texto: un tutorial en Python.

*Medium.*

<https://medium.com/datos-y-ciencia/preprocesamiento-de-datos-de-texto-un-tutorial-en-python-5db5620f1767>

GfG. (2022, 27 diciembre). *Snowball stemmer NLP*. GeeksforGeeks.

<https://www.geeksforgeeks.org/snowball-stemmer-nlp/>

DoC · *SPACY API documentation*. (s. f.). Doc. <https://spacy.io/api/doc>