Universidad de los Andes
Inteligencia de Negocios

Proyecto – Etapa 1

Análisis de textos turismo de los Alpes

Nicolás Díaz – 202021006

Sebastián Casanova– 202115116

Julio Peña– 201616539

Table of Contents

1. Entendimiento del negocio y enfoque analítico.

Oportunidad/problema Negocio	Identificar las características de una atracción turística atractiva para turistas locales e internacionales, para generar estrategias que aumenten y fomenten el turismo en Colombia. Para alcanzar este objetivo, se plantea: identificar las características destacables de las zonas turísticas, comparar las características de los sitios turísticos con alta y baja clasificación, desarrollar un mecanismo para determinar la calificación que tendrá un sitio turístico por parte de los turistas, y aplicar estrategias para identificar oportunidades de mejora en los sitios turísticos. El éxito de este proyecto se medirá mediante dos criterios. Primero, se medirá el rendimiento de los resultados por medio del f1-score, que se encargará la media armónica de la precisión y recall de los modelos. El objetivo es que se alcance un f1-score inicial de 0.50 o mayor, debido a que con esta medida está identificando correctamente un buen porcentaje de los datos relevantes en comparación con los datos irrelevantes en este problema de clasificación. Segundo, los modelos deben evidenciar las características mas relevantes y coherentes que fueron de influencia en los análisis realizados. Con esto, los stakeholders tendrán conocimiento de lo que hace que un lugar turístico tenga una buena clasificación o no, para luego ellos decidir que estrategias realizar con la información provista.
Enfoque analítico	Para abordar el objetivo de identificar las características que hacen atractivos a los sitios turísticos y desarrollar estrategias para aumentar el turismo en Colombia, se realizara un análisis de texto a las reseñas que fueron provistas por los actores de turismo. En primer lugar, se procesarán las reseñas, limpiando el texto, eliminando palabras irrelevantes, tokenizando y normalizando el formato. A continuación, se extraerán características importantes de las reseñas, como la calidad del servicio, la ubicación, la limpieza, etc., utilizando técnicas como Bag of Words (BoW) y TF-IDF. Estas características se utilizarán para entrenar los modelos de aprendizaje automático que se van a utilizar. En nuestro caso se van a utilizar los siguientes 4 modelos: un modelo de regresión línea, dos modelos de Random Forest y un modelo de Naives Bayes. Que según el contexto del problema y el uso de los algoritmos BoW y TF-IDF son los que creemos que brindaran los mejores resultados.
Organización y rol dentro de ella que se	El impacto potencial de este proyecto conllevara un beneficio para la Asociación Hotelera y Turística de Colombia – COTELCO. Dado a que se espera aumentar la competitividad de los destinos turísticos colombianos, lo

beneficia con la oportunidad definida	cual reflejaría un aumento en la demanda de alojamientos en hoteles y establecimientos turísticos que sean miembros de la asociación. A su vez, los resultados que se obtengan de proyecto proporcionarían información valiosa para los miembros de COLTECO para que se puedan tomar decisiones estratégicas sobre promoción, inversión y desarrollo de nuevos productos o servicio turísticos.
Contacto con experto externo al proyecto y detalles de la planeación	Nuestro experto externo es Mariana Gutiérrez, quien nos va a ayudar a validar el enfoque que le está dando el proyecto, a nivel de actor del sector que se verá beneficiado del proyecto, los impactos esperados y la formalidad por utilizar al momento de presentar resultados del modelo analítico construido. Lo ideal es realizar una reunión con la experta por medio de la plataforma Zoom, el día Lunes 7 de abril con la intención de empezar las actividades de la etapa 2 lo más pronto posible.

2. Entendimiento y preparación de los datos

2.1. Entendimiento

En los datos están 7875 registros, cada uno representando una review con su clasificación respectiva de algún lugar turísticos. La columna de Review son de tipo texto y, tal como dice su nombre, contiene las opiniones de distintos turistas acerca de los lugares vacacionales que visitaron. Estas opiniones pueden dar una idea de las características de estos lugares. Por otro lado, la columna Class es de tipo numérico y nos da la clasificación del 1 al 5, siendo 1 la puntuación más baja y 5 la más alta, para los lugares turísticos. Esta es nuestra variable objetivo.

Por el lado de la clasificación, los lugares turísticos que tienen una clasificación de 5 representan un 29.8% de los datos; los que tienen una clasificación de 4, un 25.0%; los que tiene 3, un 19.9%; los que tienen 2, un 14.9%; 10.3%, las que tiene clasificación de 1.

Con respecto a la completitud, los datos están excelentes. No se encuentra datos nulos, afortunadamente, lo cual facilita la preparación de los datos y se tendrá un mejor resultado en el análisis. Aun así, los datos cuentan con un 0,5% de datos repetidos, lo cual representan 38 reviews repetidas.

En términos de conteo se tuvo un mínimo de review de 33 palabras, un review de máximo de 14129 y una mediana de 217 en el conjunto de datos analizado. Contando con un promedio de 407.94 palabras por review. A su vez, se tienen un máximo promedio de 13.2 caracteres en las palabras de un review y un mínimo, en promedio, de por lo menos un carácter, en una palabra.

Para la moda, se tiene que las palabras que más aparecen son stop words, palabras que son de uso común en el lenguaje pero que no aportan mucho en el análisis, solo generan ruido. Estas palabras pueden ser: "que", "de", "y", "es", entre otras. Aun así, se puede ver que los también hay palabras relevantes para el caso que se está trabajando tal como: "servicio", "habitación",

"comida", entre otras. Este análisis se describirá más detalladamente más adelante en la preparación de los datos.

2.2. Preparación de los Datos

Para poder empezar el preprocesamiento de los datos es recomendable realizar las siguientes etapas: Limpieza de los datos, tokenizar y normalizar.

2.2.1 Limpieza de Datos

Para empezar, toca eliminar y remplazar cualquier ruido que nos puede generar información poco relevante, o conflictos en el peor de los casos, a la hora de hacer el análisis de texto, tales como:

- Palabras que no estén en ASCII (eliminar), ya que como el texto está en español puede haber tildes.
- Palabras que tengan mayúsculas (volver minúsculas), para que el programa no se confunda y tome dos palabras iguales como diferentes por tener mayúsculas.
- Puntuación (eliminar), ya que estas no se cuentan como palabras en sí y no nos daría información relevante.
- Reseñas repetidas (eliminar), para que no genere sesgos o redundancia en el análisis.
- Palabras Stop (eliminar), para eliminar palabras comunes que generan ruido en los datos y que no aportan al análisis.

2.2.2 Normalizar

Toca tomar en cuenta que hay varios aspectos de la normalización que pueden hacer parte de la limpieza de datos, y viceversa. En nuestro caso, se realizará lematización y stematizacion. Lematizar debido a que tiene en cuenta el contexto y la gramática del texto para encontrar la forma básica de una palabra. Esto ayuda a preservar el significado original de las palabras, lo que puede ser crucial para comprender las opiniones expresadas en las reseñas. Además, de que ayuda a reducir la variabilidad léxica y resaltar de manera más efectiva las características de los lugares turísticos.

Por otro lado, se stemmatiza dado a que este simplifica el texto al reducir las palabras a su forma raíz, lo que a su vez ayuda a identificar patrones. Al agrupar palabras similares bajo una misma raíz, se reduce la dimensionalidad del conjunto de datos, lo que ayuda a identificar aspectos relevantes y comunes en las reseñas. Esto permite una comprensión más clara de las características que hacen que un sitio turístico sea atractivo o no para los visitantes.

2.2.3 Tokenizar

Se realiza la tokenizacion dado a que ayuda a capturar la estructura el significado y características relevantes del texto, lo cual es nos ayuda en el contexto del negocio dado a que queremos evidenciar las características más relevantes de los lugares turísticos. Separar las frases en palabras por medio de apply y una función que tokeniza. Una vez toquenizado se aplican las funciones de limpieza de datos y normalización.

2.3 Análisis y últimos ajustes

Se va a revisar las palabras más frecuentes de los diferentes tipos de reseñas con el fin de encontrar Stopwords. Para eso se utilizará un WordCloud, el cual es una representación visual de las palabras más frecuentes en un conjunto de texto, donde el tamaño de cada palabra es proporcional a su frecuencia en el texto. Las palabras se muestran en forma de nube, donde las palabras más frecuentes suelen aparecer más grandes y destacadas. La idea es ver cuales palabras comunes, que no aportan mucho significado al texto, están en los tipos reseñas y si estas están o no repetidas entre las diferentes clasificaciones. Todo esto con el fin de mejorar la precisión de los modelos.

Se lograron identificar las siguientes palabras:

- Si
- Ser
- Cada
- Hotel
- Habit
- Mas
- Lug

Estas palabras son las que más frecuencia tenían Y que se compartían entre las clases de reseñas. Muchas de estas palabras pueden ser redundantes en el contexto del problema dado a que son lugares turísticos, además de que algunas de ellas pueden tener significado distinto.

La idea ahora es eliminar estas palabras del dataframe y dejar los datos listos para los modelos.

3. Modelado y Evaluación

Se realizaron 4 modelos: 2 Random Forest, Naives Bayes y regresión lineal. En conjunto con dos algoritmos de análisis de textos: BoW y TF-IDF.

3.1 BoW y TF-IDF

El algoritmo TF-IDF (Term Frequency-Inverse Document Frequency) es una técnica utilizada en el procesamiento de lenguaje natural (NLP) para analizar la importancia de una palabra en un documento en relación con una colección de documentos. Es comúnmente utilizado en tareas de minería de texto, recuperación de información y clasificación de texto. En el contexto del negocio, TF-IDF puede ser utilizado para identificar las palabras clave y características más relevantes que contribuyen a la evaluación de un lugar turístico.

Bag of words (BoW) es una técnica de procesamiento de texto que convierte documentos de texto en vectores numéricos, ignorando el orden de las palabras. Primero, construye un vocabulario a partir de todas las palabras diferentes en nuestro conjunto de datos. Luego, para cada review en nuestro conjunto de datos, cuenta la frecuencia de cada palabra y construye un vector donde cada elemento representa el número de ocurrencias de una palabra específica en la review. Esto resulta en una matriz donde cada fila representa una review y cada columna representa una palabra única del vocabulario.

3.2 Modelo Random Forest en conjunto con TF-IDF (Nicolas Diaz Montaña)

En este escenario, el algoritmo de Random Forest emerge como una herramienta valiosa para realizar un análisis independiente de los conjuntos de datos de reseñas de sitios turísticos. Principalmente, gracias al manejo robusto que tiene frente al sobreajuste en conjuntos de datos complejos, el Random Forest puede proporcionar información crucial sobre la importancia de diversas características en la calificación de los turistas. Esta interpretación de las características podría ayudar a identificar áreas de mejora en los sitios turísticos y diseñar estrategias efectivas que puede ser de interes para los stakeholders. Además, Random Forest ayuda con el problema de desbalance que existen en los datos. Donde existen mas reviews positivas (5 y 4) que negativas (1 y 2).

3.2.1 Validación Cuantitativa

Conjunto de Entrenamiento:

Precisión: 1.0

Recall: 1.0

F1-score: 1.0

Conjunto de Prueba:

Precisión: 0.4239

Recall: 0.4402

F1-score: 0.4147

3.3 Modelo Random Forest en conjunto en conjunto con BoW (Julio Alexander Peña Tovar)

Las razones para utilizar Random Forest en conjunto con BoW son muy similares a las que analizamos anteriormente con TF-IDF. En este contexto, lo que cambia es el algoritmo de vectorización. Random Forest resulta conveniente debido a que la matriz resultante de aplicar Bag of Words puede contener muchas palabras que no son necesariamente relevantes a la hora de relacionarlas con las calificaciones. Random Forest maneja automáticamente estos casos y nos permite obtener información más relevante, lo que hace que nuestros modelos sean más eficientes. Además, Random Forest nos proporciona información sobre la importancia de cada palabra y su impacto en las decisiones del modelo.

3.3.1 Validación Cuantitativa

Conjunto de Entrenamiento:

Precisión: 1.0

Recall: 1.0

F1-score: 1.0

Conjunto de Prueba:

Precisión: 0.4359

Recall: 0.4491

F1-score: 0.4281

3.4 Modelo Naive Bayes en conjunto con TF-IDF (Nicolas Diaz Montaña)

Naive Bayes permite clasificar las reseñas en categorías de sentimiento, mientras que TF-IDF ayuda a identificar las características únicas de cada sitio turístico a partir de las palabras utilizadas en las reseñas. Juntos, estos métodos proporcionan una visión integral de los factores que influyen en la atracción y recomendación de los sitios turísticos.

3.4.1 Validación Cuantitativa

Conjunto de Entrenamiento:

Precisión: 0.6592

Recall: 0.6146

F1-score: 0.5753

Conjunto de Prueba:

Precisión: 0.3829

Recall: 0.4107

F1-score: 0.3524

3.5 Modelo de regresión (Sebastián Casanova Ospina)

Teniendo en cuenta los objetivos que se buscan resolver para la organización, es evidente la necesidad de poder predecir la calificación que un turista le dará al lugar turístico visitado. Esto con diferentes fines que pueden beneficiar al negocio como lo es saber que lugares son preferidos por los turistas o también saber en qué poder mejorar los lugares para que los turistas tengan una mejor experiencia.

Se dice que la regresión es lineal dado que se asume una relación lineal entre las variables de entrada (features) y la variable objetivo (target). En Scikit-Learn existen varias formas de implementar el algoritmo de regresión lineal, para la elaboración de este proyecto se usará la implementación LinearRegression.

3.5.1 Validación Cuantitativa

Conjunto de Entrenamiento:

RMSE: 0.0001

R^2: 0.9999

MAE: 0.0001

Conjunto de Prueba:

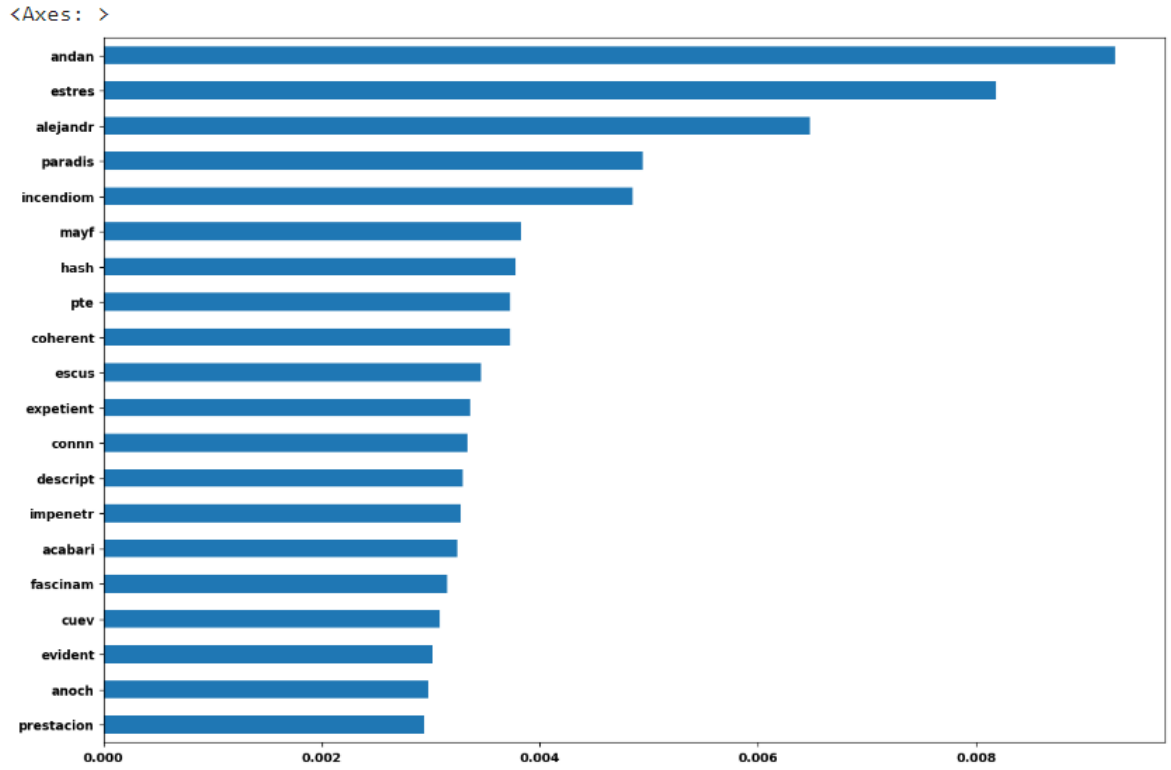
RMSE: 1.6958

R^2: -0.6763

MAE: 1.3281

4. Resultados

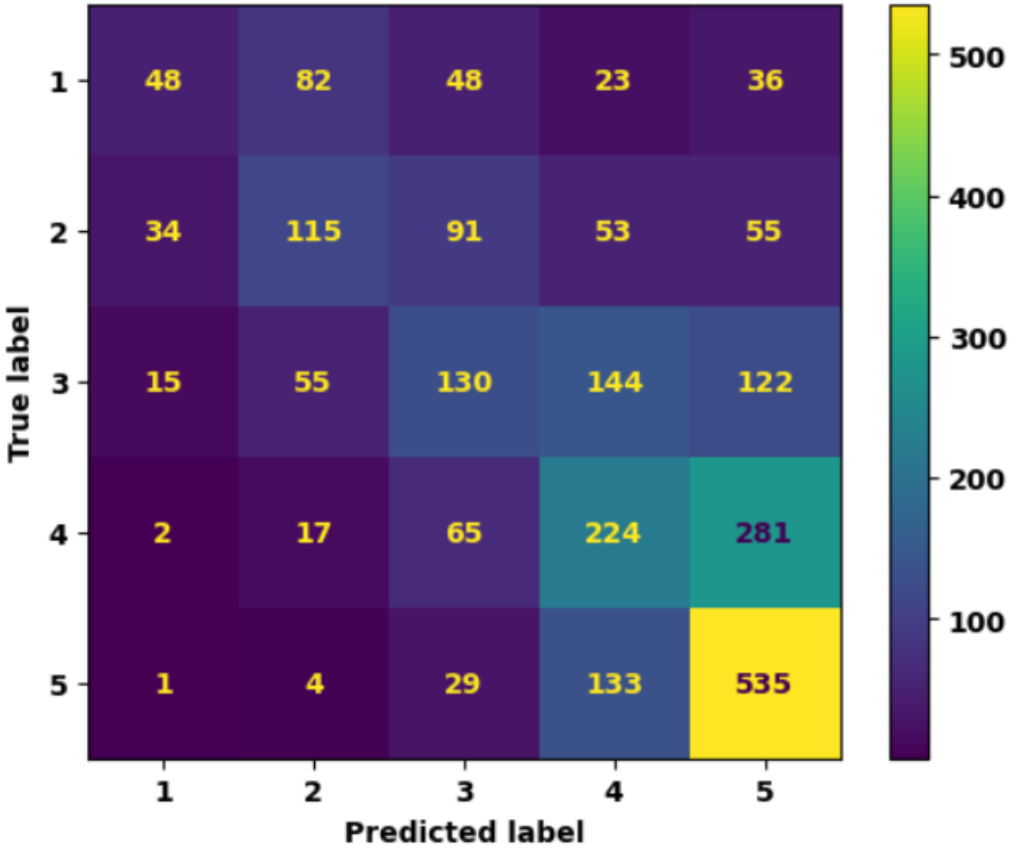
Para el análisis tomaremos como modelo guía el que tuvo los mejores resultados, para nuestro caso Random Forest Bow. A continuación, tenemos las palabras mas relevantes arrojadas en nuestro modelo, cabe aclarar que estos términos pasaron por un proceso de lematización.



Nuestro modelo es capaz de filtrar las palabras más relevantes para cada clasificación de 1-5, sin embargo, decidimos hacer más énfasis en las palabras más relevantes para el modelo en su totalidad, que se muestran anteriormente, ya que estas nos ayudan a representar los términos más clave tanto en lo positivo y lo negativo. Según las condiciones del negocio tenemos que obtener con los modelos información para aplicar estrategias para identificar oportunidades de mejora que permitan aumentar la popularidad de los sitios y fomentar el turismo. Consideramos que una palabra que pueda repetirse tanto en reviews positivas como negativas deberían ser a las que más debería ponerse atención ya que su impacto dependiendo de su ejecución puede ser positivo o negativo, lo cual le da más relevancia que ítems que solo pueden ser por ejemplo positivos, pero que cuya ausencia no genera consecuencias.

Las palabras y sus derivados entonces son tanto los puntos que más representan la diferencia entre una buena o mala experiencia y recomendamos poner énfasis en los temas relacionados a estas para lograr el objetivo de mejora, aumento de popularidad y el turismo.

```
Out[134]:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x21290a482b0>
```



5.Mapa de actores relacionado con el producto de datos creado

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Equipo de Análisis de Datos	Equipo Interno	Mayor comprensión de las características que hacen atractivos los destinos turísticos, lo que facilita la toma de decisiones estratégicas y la creación de campañas de promoción más efectivas.	Riesgo de que el modelo no proporcione insights relevantes o precisos, lo que podría resultar en la toma de decisiones erróneas y desperdicio de recursos.
Equipo de Marketing y Promoción	Equipo Interno	Acceso a información detallada sobre las preferencias y comportamientos de	Posibilidad de malinterpretar los resultados del modelo y basar campañas de

		los turistas, lo que permite diseñar estrategias de marketing más segmentadas y personalizadas para promover los destinos turísticos de manera efectiva.	marketing en información incorrecta, lo que podría afectar la reputación de la organización y la atracción de turistas.
Equipo de Desarrollo de Productos y Servicios	Equipo Interno	Mejora en la oferta de productos y servicios turísticos, al contar con insights basados en datos sobre las preferencias y necesidades de los turistas, lo que aumenta la satisfacción del cliente y la competitividad de la organización.	Riesgo de depender demasiado del modelo y descuidar otros aspectos importantes de la gestión de productos y servicios, como la calidad y la innovación, lo que podría afectar la experiencia del cliente.
Alta Dirección	Gestión	Mecanismo de toma de decisiones basado en información objetiva y precisa, lo que facilita la planificación estratégica y el cumplimiento de los objetivos organizacionales relacionados con el turismo.	Posibilidad de tomar decisiones erróneas si la información proporcionada por el modelo no es confiable o relevante, lo que podría afectar la reputación y el desempeño general de la organización.

6. Trabajo en equipo

Roles:

- Líder del Proyecto: Nicolás Díaz.
  El líder del proyecto se encargó de la división del trabajo y la toma de decisiones. Además, estuvo al tanto del progreso de los demás miembros y supervisó cada una de las tareas que se llevaron a cabo.
- Líder de datos: Nicolas Diaz.
  El líder de datos gestiono los datos que se usaron para la construcción de los diferentes modelos y dejo disponibles todos los datos para que los demás miembros los usaran.
- Líder de Negocio: Sebastián Casanova.

El líder del negocio fue responsable de la identificación y análisis del problema de negocio que debíamos afrontar y para el cual se propuso el proyecto.

- Líder de Analítica: Julio Peña.
  El líder de analítica verifico que los entregables cumplen con los estándares de análisis y que se propuso el mejor modelo posible para las condiciones en las que se manejó el proyecto.

Algoritmo y número de horas dedicado a las tareas:

- Nicolás Díaz: Modelo Random Forest en conjunto con TF-IDF y Modelo Naive Bayes en conjunto con TF-IDF. 14 horas.
- Julio Peña: Bag of words. 9 horas.
- Sebastián Casanova: Regresión Lineal 10 horas.

Distribución de puntos:

- Nicolás Díaz: 40
- Julio Peña: 30
- Sebastián Casanova: 30