Proyecto: Turismo de los Alpes

1. Equipo de trabajo y tareas asignadas 2
2. Entendimiento del negocio y enfoque analítico 2
   2.1. Objetivos y criterios de éxito 2
   2.2. Enfoque analítico 3
   2.3. Contacto con equipo externo 4
3. Entendimiento y preparación de los datos 4
   3.1. Entendimiento de los datos 4
   3.2. Calidad de los datos 5
   3.3. Preparación de los datos 5
4. Modelado y evaluación 6
   4.1. Support Vector Machines (Pablo Pedreros Díaz) 6
   4.2. Regresión Ridge (Nicolás Camargo) 6
   4.3. Random Forest (Jairo Garavito Correa) 6
5. Resultados 7
   5.1. Support Vector Machines (Pablo Pedreros Díaz) 7
   5.2. Regresión Ridge (Nicolás Camargo) 8
   5.3. Random Forest (Jairo Garavito Correa) 8
6. Mapa de actores 9
7. Conclusiones 9

1. Equipo de trabajo y tareas asignadas

El equipo de trabajo está conformado por 3 integrantes:

- Nicolas Camargo: Líder de analítica. Encargado de la implementación de dos de los modelos utilizados. Encargado de evaluar los análisis cualitativos y cuantitativos de los modelos y definir los criterios de éxito de estos, de proponer posibles mejoras en estos y de hacer la selección del modelo final documentado todo su trabajo. 7 horas de trabajo.
- Pablo Pedreros: Líder de proyecto y líder de datos. Encargado de la planeación de las reuniones del proyecto, de la asignación de tareas en el grupo y de la entrega. Como líder de datos, encargado de la selección de los datos, del perfilamiento y análisis de estos y de proponer y evaluar la limpieza y preparación de estos. Adicionalmente, encargado de realizar dos de los modelos usados en el proyecto y documentar estas tareas. 10 horas de trabajo.
- Jairo Garavito Correa: Líder de negocio. Encargado de definir los actores afectados y los objetivos del proyecto. Encargado del análisis del modelo final de cara al negocio y concluir sobre cómo los modelos resuelven o no la problemática y comunicar la información encontrada de forma clara para el negocio. Encargado de la implementación de uno de los modelos propuestos. Documentar todo su trabajo. 8 horas de trabajo.

El principal reto del proyecto fue implementar un modelo con las métricas de calidad requeridas, pues al tratar textos subjetivos como las reseñas los modelos tienden a clasificar erróneamente los textos similares en categorías similares, por lo que hubo que probar varios tipos de modelos intentando hacer una clasificación estricta, pero sin llegar al sobreajuste.

Según lo trabajado por el grupo, los 100 puntos serían repartidos equitativamente. Para próximos proyectos la única mejora sería hacer contacto más estrecho y temprano con el equipo de expertos para enfocar el proyecto correctamente desde el inicio.

2. Entendimiento del negocio y enfoque analítico

En base a la información suministrada para el proyecto se realizó el siguiente análisis sobre el negocio.

2.1. Objetivos y criterios de éxito

El proyecto tiene como objetivo primordial analizar las características distintivas de los destinos turísticos en Colombia que atraen tanto a visitantes locales como extranjeros. Se busca comprender en detalle qué aspectos hacen que ciertos lugares sean altamente recomendados por los turistas, en contraste con aquellos que reciben bajas calificaciones. Esto permitirá identificar áreas de mejora y desarrollar estrategias efectivas para aumentar la popularidad y el atractivo de los sitios turísticos menos favorecidos.

Un aspecto crucial del proyecto es la creación de un mecanismo de calificación de los sitios turísticos basado en las reseñas de los visitantes. Esto no solo proporcionará una evaluación objetiva de la satisfacción del turista, sino que también permitirá tomar decisiones informadas para la mejora continua de los destinos turísticos en Colombia con base en las características determinantes encontradas. Además, se espera que la implementación de estrategias basadas en estos análisis conduzca a un aumento en el número de turistas que visitan el país, lo que a su vez tendrá un impacto positivo en la economía nacional y en la generación de empleo, especialmente en sectores relacionados con el turismo, como la hotelería y el transporte.

El proyecto no solo busca impulsar la economía colombiana a través del turismo, sino también promover la diversidad cultural y natural del país. Al resaltar los atractivos turísticos de Colombia, se espera que se fomente la conservación del patrimonio cultural y medioambiental, contribuyendo así a la valoración y preservación de la riqueza cultural y natural del país. Además, al mejorar la percepción internacional sobre Colombia como destino turístico seguro y atractivo, se espera que el proyecto tenga un impacto positivo en la imagen del país a nivel nacional e internacional.

Definimos como criterios de éxito un análisis cuantitativo que incluya métricas de precisión y sensibilidad que demuestren un modelo funcional con utilidad para categorización de las reseñas de los usuarios y un análisis cualitativo que permita ver las características más relevantes de una buena reseña, así como las principales características de una reseña negativa, que definen los aspectos a tener en cuenta al tomar decisiones para un producto turístico

2.2. Enfoque analítico

Para alcanzar los objetivos del negocio de manera analítica, el enfoque propuesto se basa en la construcción de un modelo de análisis de textos que permita calificar automáticamente nuevas reseñas de sitios turísticos con un alto nivel de precisión y sensibilidad. Este enfoque se divide en varias etapas:

1. **Preparación de datos:** Se limpiarán los datos de reseñas de sitios turísticos junto con sus calificaciones asociadas. Es importante asegurarse de que los datos estén completos, sean coherentes y estén libres de ruido o datos irrelevantes.
2. **Selección y representación de características:** Se seleccionarán las palabras o características más relevantes de las reseñas que puedan influir en la calificación de los sitios turísticos. Estas palabras se utilizarán para representar las reseñas como variables en el modelo analítico. Se van a utilizar técnicas como la tokenización, eliminación de palabras irrelevantes (stopwords) y lematización para procesar el texto y extraer las características más importantes.
3. **Desarrollo del modelo analítico:** Se desarrollará un modelo de procesamiento de textos. Para esto se van a explorar diferentes algoritmos,

como Multinomial Naive Bayes, árboles de decisión o Support Vector Machines, para construir un modelo que sea capaz de predecir la calificación de las reseñas basándose en las palabras seleccionadas como características.

4. **Validación del modelo:** Se evaluará el rendimiento del modelo utilizando técnicas de validación cruzada y métricas como precisión, sensibilidad (recall), precisión, y F1-score. Es crucial asegurarse de que el modelo sea capaz de generalizar correctamente a nuevos datos y que tenga un rendimiento satisfactorio en la clasificación de reseñas.

2.3. Contacto con equipo externo

El 6 de abril se realizó el encuentro con los expertos del curso de estadística para que aprobaran el enfoque analítico del proyecto.

3. Entendimiento y preparación de los datos

3.1. Entendimiento de los datos

Para este proyecto, el negocio brindó 2 archivos de formato CSV, uno para realizar el entrenamiento de los modelos y el otro para predecir las calificaciones. Los atributos que tiene el primer archivo son:

- Reviews: Objeto de tipo String que tiene las reseñas realizadas por los clientes sobre los distintos sitios turísticos.
- Class: Un número entero entre 1 y 5 que representa la calificación dada por el cliente para el sitio turístico.

El archivo cuenta con un total de 7875 registros.

En cuanto a la distribución de las calificaciones se pueden observar las siguientes distribuciones:

Calificación	Total
1	778
2	1185
3	1574
4	1977
5	2361

Se puede observar como la calificación que más se repite es 5, y mientras más baja es la calificación hay un menor número de calificaciones.

En cuanto a las reviews, se obtuvieron las siguientes estadísticas:

Max length	10419
Median length	1543
Mean length	408.40495
Min length	9
Total characters	3216189

Distinct characters	188

Además de esto, el 98% de los caracteres pertenecen al código de caracteres ASCII, mientras que el 2% restante se distribuye entre signos de puntuación, caracteres especiales como las vocales con tilde, y emojis. Los caracteres pertenecen todos al latín o son comunes, por lo que no deben existir problemas para manejar los tipos de caracteres.

Por último, se puede observar que las palabras que más se repiten son conectores como preposiciones y conjunciones, que por sí solas no brindan información relevante para el modelo.

3.2. Calidad de los datos

En cuanto a la calidad de los datos, se obtuvieron los siguientes resultados:

- **Completitud:** No existen datos faltantes. Todas las clases cuentan con los 7875 registros que tiene el archivo.
- **Unicidad:** El documento cuenta con 73 registros completamente repetidos, es decir, hay 7802 registros únicos en los datos.
- **Consistencia:** Por un lado, la variable “reviews” al ser un texto no tiene muchas restricciones del formato, por lo que no existen datos inconsistentes. Por otra parte, en la variable “class” todos los registros son enteros entre 1 y 5, por lo que todos los datos son consistentes.
- **Validez:** Aunque haya registros repetidos, estos son muy pocos, por lo que los datos siguen siendo válidos y significativos para proyecto.

3.3. Preparación de los datos

La limpieza realizada consiste en los 4 siguientes pasos:

- **Limpieza de datos:** Se inicia eliminando registros duplicados y se realiza una transformación de todos los caracteres a minúsculas para uniformizar el análisis. Luego, se elimina la puntuación y las stopwords, palabras vacías en español que no aportan significativamente al análisis.
- **Tokenización:** Se divide cada review en palabras individuales (tokens) para facilitar la clasificación posterior. Esto permite un análisis de sentimiento basado en el contenido de cada reseña.
- **Normalización:** Se aplica la técnica de stemming y lematización para simplificar las palabras y mejorar su comprensión y utilidad en el modelado. El stemming reduce las palabras a su forma raíz, mientras que la lematización las reduce a su forma base.
- **Selección de campos:** Se elige la variable objetivo, en este caso la calificación otorgada por los clientes, y se aplican técnicas de vectorización al texto, incluyendo binarización, conteo de palabras y TF-IDF (vectorización basada en el conteo de palabras, pero reduciendo la importancia de palabras usadas a lo largo de muchos documentos), para determinar cuál técnica produce los mejores resultados en la clasificación de las reseñas.

4. Modelado y evaluación

Para este proyecto se diseñaron 5 modelos predictivos. Todos los algoritmos se encuentran explicados e implementados en el notebook. De estos modelos se tomó la decisión de presentar los 3 más importantes en este documento.

4.1. Support Vector Machines (Pablo Pedreros Díaz)

Para nuestro enfoque en el análisis de sentimientos, consideramos utilizar Support Vector Machines (SVM). Estos algoritmos son especialmente adecuados para manejar espacios de variables con muchas dimensiones, lo cual es común en problemas de análisis de texto donde cada palabra puede considerarse una dimensión. SVM calcula un hiperplano que separa los puntos en diferentes grupos, basándose en los datos de entrada. Este hiperplano se calcula de manera que maximiza su distancia a los puntos más cercanos de las clases, lo que se conoce como el margen máximo.

SVM es útil en el análisis de sentimientos porque permite encontrar un límite de decisión óptimo entre las diferentes clases de sentimientos presentes en las reseñas. Su implementación implica la optimización de un problema de programación cuadrática para encontrar el hiperplano óptimo de separación. Además, SVM es efectivo en la identificación de patrones complejos en los datos y puede manejar eficazmente conjuntos de datos grandes, lo que lo convierte en una opción prometedora para nuestro proyecto de análisis de sentimientos.

4.2. Regresión Ridge (Nicolás Camargo)

La regresión Ridge es una técnica de regresión utilizada para predecir valores numéricos, en contraposición a los algoritmos de clasificación que predicen clases. A diferencia de la regresión lineal estándar, la regresión Ridge incorpora un término de regularización que ayuda a prevenir el sobreajuste cuando hay muchas variables de entrada. Este término de regularización penaliza los coeficientes grandes, lo que lleva a coeficientes más pequeños y estables.

Este algoritmo calculará un coeficiente para cada variable de entrada para estimar un valor de la calificación. Como este valor que estima será un número flotante y no un número entero, para la predicción de los valores redondearemos los resultados que nos haya dado la regresión para que queden en números enteros, respetando el límite máximo de 5 y el mínimo de 1.

La regresión Ridge puede ser una opción prometedora para nuestro problema, ya que su capacidad para controlar el sobreajuste es particularmente útil cuando se trabaja con conjuntos de datos complejos y numerosas variables predictoras.

4.3. Random Forest (Jairo Garavito Correa)

Random Forest es un algoritmo de aprendizaje supervisado utilizado para tareas de clasificación y regresión. Pertenece a la categoría de algoritmos de ensemble, lo

que significa que combina múltiples modelos de aprendizaje para mejorar la precisión y la generalización del modelo final.

El funcionamiento de Random Forest se basa en la construcción de un gran número de árboles de decisión durante el proceso de entrenamiento. Cada árbol de decisión se construye utilizando un subconjunto aleatorio de características del conjunto de datos y un subconjunto aleatorio de ejemplos de entrenamiento (bootstrap sampling). Luego, durante la fase de predicción, cada árbol en el bosque emite una predicción y la clase o valor final se determina por votación (en el caso de clasificación) o por promedio (en el caso de regresión) de las predicciones individuales de los árboles.

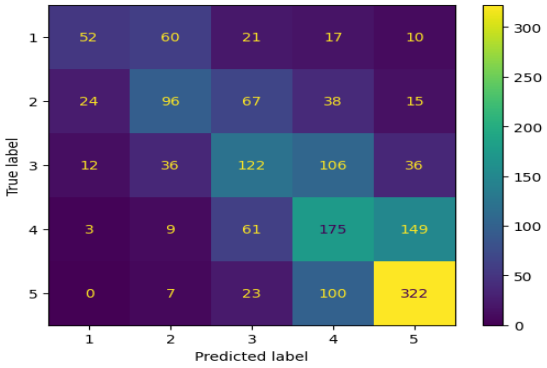
Random Forest es una opción sólida para el proyecto de análisis de sentimientos debido a su robustez ante datos ruidosos y atípicos, lo que lo hace adecuado para conjuntos de datos de reseñas diversos. Además, su capacidad para manejar automáticamente la selección de características elimina la necesidad de una selección manual. Al construir múltiples árboles de decisión y promediar sus predicciones, Random Forest reduce el riesgo de sobreajuste, lo que lo hace menos propenso a capturar patrones espurios en los datos de entrenamiento. Aunque no proporciona una interpretación directa de los factores que influyen en las predicciones, es posible evaluar la importancia relativa de las características utilizando métricas específicas.

5. Resultados

5.1. Support Vector Machines (Pablo Pedreros Díaz)

Estadística	**Resultado**
Train Accuracy	1
Test Accuracy	0.49
Precision	0.49
F1-score	0.48

Matriz de confusión:



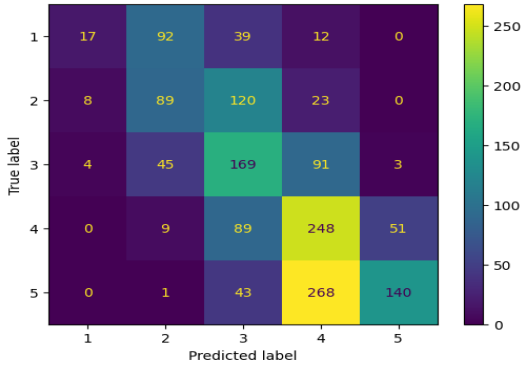
Se puede ver que los valores de precisión y de F1 aumentaron ligeramente frente a los algoritmos anteriores, llegando casi a 0.5. Igualmente, el algoritmo sigue cometiendo los mismos errores de los algoritmos anteriores de asignar las reseñas

a calificaciones cercanas a la real. A pesar de esto, el algoritmo sería útil para tener una primera idea de la calificación de una reseña pues tiende a seguir los patrones que hacen que una calificación sea más alta a pesar de no ser estrictamente preciso, probablemente por patrones muy similares entre las reseñas de diferentes calificaciones.

5.2. Regresión Ridge (Nicolás Camargo)

Estadística	**Resultado**
Test Accuracy	0.42
Precision	0.50
F1-score	0.41

Matriz de confusión:

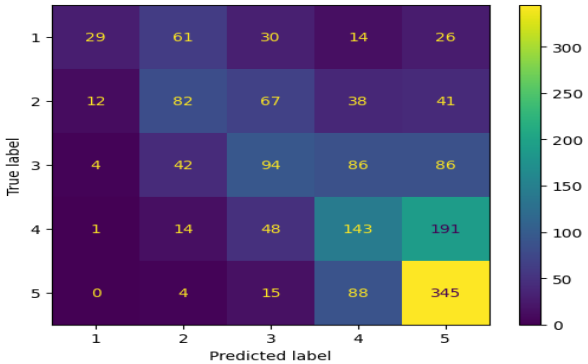


Vemos que nuevamente hubo un gran porcentaje de error en las predicciones, especialmente en valores de calificación 4 que se llevaron a calificación 5, por lo que este modelo sigue teniendo unas métricas un poco peores que las de la Support Vector Machine.

5.3. Random Forest (Jairo Garavito Correa)

Estadística	**Resultado**
Train Precision	1
Test Precision	0.44
Accuracy	0.44
F1-score	0.42

Matriz de confusión:



Los resultados obtenidos para el modelo de Random Forest muestran una discrepancia notable entre la exactitud en el conjunto de entrenamiento y la exactitud en el conjunto de prueba. Esta discrepancia sugiere que el modelo puede estar sobreajustado a los datos de entrenamiento y no generaliza bien a nuevos datos. La precisión y la puntuación F1 son relativamente bajas, lo que indica que el modelo tiene dificultades para clasificar correctamente las reseñas en todas las categorías de sentimiento.

6. Mapa de actores

En la identificación de los actores se construyó la siguiente tabla:

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Turista	Beneficiado	Acceso a reseñas detalladas para decisiones de viaje más informadas.	Dependencia excesiva de opiniones individuales que podrían no coincidir con las expectativas personales.
Sitios turísticos	Cliente	Identificación de áreas de mejora y aumento de la popularidad gracias al análisis de las reseñas.	Riesgo de críticas negativas públicas que podrían afectar la reputación del sitio si no se gestionan adecuadamente.
Ministerio de Comercio, Industria y Turismo de Colombia	Financiador	Mejora del turismo y desarrollo económico a través de la promoción efectiva de destinos turísticos.	Riesgo de inversión no justificada si los resultados del análisis no se traducen en mejoras tangibles en el sector turismo.
Asociación Hotelera y Turística de Colombia – COTELCO	Proveedor de datos	Adaptación de servicios y promociones para satisfacer mejor las necesidades de los clientes.	Riesgo de violación de privacidad de datos si no se manejan adecuadamente.

7. Conclusiones

Para el modelo final se escogió el modelo de Support Vector Machines que es el que, por un muy pequeño margen, obtuvo mejores métricas. El modelo no obtuvo unos valores idóneos en el análisis cuantitativo, pues el modelo frecuentemente clasifica reseñas de una clase como si fueran de una clase vecina. A pesar de esto, el modelo es congruente con la realidad de los datos y sus predicciones sí dan cuenta, con suficiente precisión, de qué reseñas son positivas y cuáles son negativas, pues, aunque confunda reseñas de cierta clase con clases vecinas,

prácticamente no confunde reseñas totalmente positivas con totalmente negativas. Un análisis de los resultados de este modelo puede ayudar a los atractivos turísticos a encontrar las características más relevantes para los clientes al escribir una reseña.

Entonces hacemos las predicciones para el set de datos sin clasificación y creamos una nube de palabras para las reseñas de calificación 1.

Para la nube de palabras de la clase 1, igual que con la 5, muchas de las palabras repetidas con frecuencia no aportan al análisis, pero hay otras que nos pueden dar información relevante. Podemos ver que “comida” y “restaurante” fueron de los conceptos más relevantes entre los datos, por lo que podemos pensar que hubo un descontento general con los servicios de alimentación de los sitios turísticos considerados. Además, podemos ver “habitación”, “cama”, “baño”, “sucia” y, en menor medida, “ruido” y “ducha”, lo que nos puede indicar que las condiciones de las habitaciones fueron adecuadas para muchos hoteles. Por último, vemos que “pagamos” y “servicio” también fueron relevantes, por lo que inferimos que para los clientes la relación calidad/precio de los servicios no fue adecuada y que el servicio del personal, sea de restaurante o del hotel tampoco lo fue.

Para la clasificación de la clase 5, podemos ver que estas reseñas estuvieron especialmente relacionadas con un excelente servicio por parte del personal. Además, destacan “ciudad” y “vista”, lo que nos indica que probablemente los hoteles en zonas más relevantes de la ciudad y con buenas vistas fueron favoritos para los clientes. Por último, vemos un fuerte énfasis en “comida” y “restaurante”, lo que nos dice que un buen servicio de alimentación fue diferencial para los clientes.

Con estas conclusiones que otorga el modelo, los servicios deberían estar en capacidad de analizar los conceptos relevantes y tomar mejores decisiones sobre en qué áreas invertir recursos para mejora de los servicios, así como priorizar a la hora de crear nuevos hoteles.