**Proyecto interdisciplinario analítica de textos**

**Inteligencia de negocios**
**Estadística**


# Turismo de los Alpes

**Isabella Nova**
**Felipe Rueda**
**Santiago Pardo**
**Luis Felipe Plazas**
**Ana Sanchez**

**06 de abril de 2024**
**Bogotá D.C**

## Tabla de contenidos

1. **Contexto:**

El Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia – COTELCO, cadenas hoteleras de la talla de Hilton, Hoteles Estelar, Holiday Inn y hoteles pequeños ubicados en diferentes municipios de Colombia están interesados en analizar las características de sitios turísticos que los hacen atractivos para turistas locales o de otros países, ya sea para ir a conocerlos o recomendarlos.

De igual manera, quieren comparar las características de dichos sitios, con aquellos que han obtenido bajas recomendaciones y que están afectando el número de turistas que llegan a ellos. Adicionalmente, quieren tener un mecanismo para determinar la calificación que tendrá un sitio por parte de los turistas y así, por ejemplo, aplicar estrategias para identificar oportunidades de mejora que permitan aumentar la popularidad de los sitios y fomentar el turismo.

Esos actores de turismo prepararon dos conjuntos de datos con reseñas de sitios turísticos. Cada reseña tiene una calificación según el sentimiento que tuvo el turista al visitarlo. Estos actores quieren lograr un análisis independiente de los conjuntos de datos y al final del proyecto discutir sobre los grupos de científicos de datos e ingenieros de datos que acompañarán el desarrollo real de este proyecto.

| **Título** | **Descripción** |
|---|---|
| **Oportunidad/problema Negocio** | Identificar y mejorar las características de sitios turísticos para aumentar su atractivo y recomendaciones, impactando positivamente en el flujo turístico. |
| **Enfoque analítico** | Se utilizarán técnicas de procesamiento de lenguaje natural (NLP) para analizar las reseñas y asignar calificaciones de sentimiento. Algoritmos como análisis de sentimiento, clasificación y clustering serán aplicados para comparar sitios y detectar patrones. |

| | |
|---|---|
| **Organización y rol dentro de ella que se beneficia con la oportunidad definida** | El Ministerio de Comercio, Industria y Turismo y asociaciones como COTELCO, así como cadenas hoteleras y hoteles pequeños, se beneficiarán como promotores y directos interesados en la mejora del turismo. |
| **Contacto con experto externo al proyecto y detalles de la planeación** | Se coordinará con expertos en estadística y análisis de datos para revisar los resultados preliminares y planificar la etapa 2 del proyecto, asegurando un enfoque objetivo y especializado. |

## 2. Entendimiento del negocio y enfoque analítico

### 2.1 Objetivos del proyecto:

El proyecto descrito anteriormente, está propuesto para que se cumplan unos objetivos concretos, y así poder generar soluciones para el sector del turismo. Uno de estos es analizar las características de los sitios turísticos, que los hacen atractivos tanto para los turistas locales como extranjeros. De esta manera, se pueden mejorar estas particularidades, y los hoteles sean más visitados y mejor calificados. Esto ayudará tanto a la comunidad local a que su zona esté más visitada, como a los turistas para que puedan tener una mejor experiencia.

Por otro lado, se quiere comparar las características de los sitios turísticos populares, con aquellos que han tenido bajas recomendaciones. Con lo anterior, se puede realizar un análisis de cuales son la razones por las cuales los sitios peor calificados han obtenido estas notas, y generar un plan de acción para mejorarlos. Esto va a ayudar al sector público o privado que esté invirtiendo en estos sitios, y se puedan guiar en hacia dónde centrar sus inversiones y esfuerzos.

Es vital poder desarrollar un mecanismo para predecir la calificación que un sitio recibirá de los turistas, según los comentarios que este haya tenido. Este es el objetivo principal del proyecto, y es esencial para que todos los actores del turismo en Colombia se vean beneficiados. El poder identificar puntos

clave de cómo mejorar estos sitios turísticos, va a colaborar a toda la comunidad local, a los turistas y a los sectores que están invirtiendo dentro de la  hotelería y turismo en Colombia

**2.2 Criterios de éxito:**

Es fundamental contar con indicadores de medición y éxito al finalizar un proyecto de esta magnitud, ya que estos proporcionan una medida clara y objetiva para evaluar si se cumplieron los objetivos y metas establecidos anteriormente. Al momento de finalizar el proyecto, se pudo identificar precisamente las características que hacen atractivos a los sitios turísticos. Esta precisión será medida con diferentes indicadores al momento de realizar los modelos, que nos podrán mostrar qué tan preciso es este.

También se pudo comparar con precisión los sitios turísticos populares, y los menos populares, calculando las calificaciones de estos teniendo en cuenta los comentarios realizados por turistas. El medidor que se tiene para esto es el f1 score, que es una métrica utilizada comúnmente en problemas de clasificación para evaluar el rendimiento de un modelo. Se calcula como la media armónica de la precisión y la sensibilidad (recall) del modelo. La precisión mide la proporción de predicciones positivas correctas entre todas las predicciones positivas, mientras que la sensibilidad (recall) mide la proporción de instancias positivas que el modelo identifica correctamente. Se espera que este esté alrededor de 0.5, ya que un f1 score cercano a 1, significa que el modelo está sobre ajustado, y que este no será útil para el sector de hotelería y turismo.

Es vital poder identificar efectivamente las oportunidades para los sitios turísticas, es por esto que para medir ello al final del proyecto, se se utilizará una combinación de métricas como el F1 score, que proporciona una evaluación equilibrada del rendimiento del modelo de clasificación, junto con otras métricas específicas del negocio, como la precisión y la sensibilidad. Estos criterios de éxito garantizan que se haya logrado el objetivo principal del proyecto: desarrollar un mecanismo efectivo para predecir la calificación de los sitios turísticos según los comentarios de los turistas. Además, la capacidad para identificar oportunidades de mejora para los sitios turísticos contribuirá al crecimiento y desarrollo del sector turístico en Colombia, beneficiando tanto a la comunidad local como a los inversionistas en la industria hotelera y turística.

**2.3 Enfoque analítico:**

El enfoque analítico propuesto para alcanzar los objetivos del negocio se compone de varias etapas fundamentales. En primer lugar, se realizará un análisis descriptivo exhaustivo de los datos recopilados sobre los sitios turísticos. Este análisis permitirá comprender en profundidad las características y atributos de los destinos, así como identificar patrones y tendencias que puedan influir en su popularidad y calificación por parte de los turistas.

Posteriormente, se llevará a cabo un análisis comparativo entre los sitios turísticos populares y los menos populares. Esta etapa implica la identificación de diferencias significativas en términos de características, servicios ofrecidos, ubicación geográfica, y otros factores relevantes. Esta comparación ayudará a entender qué aspectos específicos hacen que algunos destinos sean más atractivos que otros, proporcionando información valiosa para la toma de decisiones estratégicas.

Luego, se aplicarán técnicas de modelado predictivo, como regresión, máquinas de vectores de soporte o redes neuronales, para predecir las calificaciones de los sitios turísticos. Estos modelos utilizan variables relevantes identificadas durante el análisis descriptivo y comparativo para prever las calificaciones que podrían recibir los destinos turísticos. Esto permitirá anticipar el potencial éxito o fracaso de un sitio en función de sus características específicas, facilitando la planificación y la toma de decisiones en el sector turístico.

Finalmente, se realizará un análisis de oportunidades de mejora utilizando técnicas como el análisis de causa raíz o el análisis FODA. Este análisis profundizará en las fortalezas, debilidades, oportunidades y amenazas de cada sitio turístico, con el objetivo de identificar áreas de mejora y potenciales estrategias para incrementar su atractivo y popularidad entre los turistas. Este enfoque holístico permitirá no solo comprender los factores que influyen en el éxito de los destinos turísticos, sino también desarrollar planes de acción concretos para impulsar el crecimiento y la competitividad en la industria del turismo.

### 2.4 Planeación

Las estudiantes del curso de estadística con las cuales vamos a trabajar para este proyecto son **Isabella Nova** y **Ana Sánchez**. Con ellas han habido varias reuniones a lo largo del proyecto, y la planeación y seguimiento de resultados de este.

Los actores del sector turístico que se verán beneficiados incluyen a las agencias de viajes, empresas hoteleras, proveedores de servicios turísticos,

autoridades locales y regionales de turismo, así como a los propios turistas. Todos estos actores podrán utilizar los resultados del proyecto para comprender mejor las preferencias de los turistas, mejorar la oferta turística, planificar estrategias de marketing más efectivas y tomar decisiones informadas para promover el desarrollo sostenible del sector turístico en Colombia, y mejorar los sitios actuales que se encuentran dentro del país

Se espera que los impactos del proyecto sean significativos en términos de mejora de la experiencia del turista, aumento de la competitividad de los destinos turísticos, incremento de la llegada de turistas tanto nacionales como internacionales, y fortalecimiento de la economía local a través del turismo. Además, se anticipa una mayor eficiencia en la asignación de recursos y una mejor planificación del desarrollo turístico a nivel local y regional.

Los resultados del modelo analítico serán presentados de manera formal a través de un informe detallado que incluirá una descripción de los objetivos propuestos, metodología utilizada, los hallazgos clave, resultados, las conclusiones y recomendaciones para los diferentes actores del sector turístico. También se tendrá el documento fuente de donde se encuentran los modelos implementados, con todas las métricas y resultados. También se entregará un video donde se realizará el análisis de los resultados, y cómo estos serán de suma importancia para la hotelería y turismo en Colombia.

La reunión para presentar los resultados de la Etapa 1 y dar inicio a la Etapa 2 está programada para el día 10 de abril. Se llevará a cabo a través de la plataforma Zoom. Durante esta reunión, se compartirán los hallazgos y análisis preliminares, se discutirán posibles ajustes o ampliaciones del proyecto, y se establecerán los objetivos y acciones para la siguiente fase del trabajo.

## 3. Entendimiento y preparación de los datos.

Primero, se recopiló toda la información que se dio de opiniones de turistas sobre diferentes lugares turísticos. Cada opinión tenía una calificación que reflejaba cuánto le gustó al turista el lugar.

Luego, de analizar las reseñas de los turistas, se sacaron las siguientes conclusiones:

- Hay 23 reseñas o datos duplicados, por lo que es necesario eliminarlos para garantizar un mejor rendimiento en el modelo pues es información repetida.

- En total, se utilizarán 6300 datos con dos características la reseña y su puntuación
- El porcentaje de datos diferentes es de 99.6%, y se debe principalmente a la cantidad de datos duplicados
- No hay datos únicos, por lo que el índice de completitud debería ser de 100

Después, se aseguró de que todas las opiniones estuvieran escritas de la misma manera. Se eliminaron los caracteres que no pertenecían al código ASCII (como varios puntos o símbolos), se convirtieron todas las palabras a minúsculas, se quitaron los signos de puntuación (incluyendo tildes) y las palabras que no aportan mucha información (como “el”, “la”, “y”, “o”), y se hizo el reemplazo de los números que estaban escritos como dígitos en letras, es decir, por ejemplo los 1, 2 y 3 pasaron a ser uno, dos y tres.

Finalmente, se dividió cada opinión en palabras individuales y se simplificó a su forma básica. Por ejemplo, las palabras “corriendo”, “correr” y “corrió” las simplificamos todas a “correr”. Esto ayuda a entender mejor el significado general de cada opinión.

Todo este proceso ayuda a tener la base para implementar diferentes modelos de procesamiento de lenguaje natural que ayuden posteriormente a entender mejor las opiniones de los turistas y a descubrir qué características hacen que un lugar turístico sea atractivo. Con esta información, se puede ayudar a mejorar los lugares que no son tan populares y a promover el turismo en Colombia.

## 4. Modelado y evaluación.

Para el modelado de esta etapa, en primer lugar se consideró el desarrollo de un pipeline para hacer la prueba de todos los modelos a realizar, para ello fue necesario tener en cuenta todos los pasos que se hicieron en la etapa de limpieza hasta el proceso de vectorización de las palabras, se realizó un pipeline para el procesamiento de datos con el objetivo de reducir el tiempo de implementación, en concreto para los datos de prueba que serán introducidos al modelo.

El primer modelo utilizado fue un modelo de Naive Bayes, el cual es uno de los algoritmos de Machine Learning más utilizados para clasificar y predecir una clase de un conjunto de datos, basado en seguir las características de una distribución de probabilidad, y siendo la más popular para problemas de clasificación de texto (como este problema). Existen varios tipos de modelo como lo son el Gaussiano, el multinomial y el de bernoulli, los cuales, al igual que con Support Vector Machines, serán probados empíricamente cuál modelo es mejor con las pruebas. Por lo tanto, para esta selección de

modelo, se intentará probar 2 modelos, los cuales son el multinomial y la distribución de Bernoulli y para el mejor modelo se escogerá aquel que tenga mejores métricas en prueba

El segundo modelo utilizado fue una regresión ridge, la cual consiste en regularizar el modelo resultante imponiendo una penalización al tamaño de los coeficientes de la relación lineal entre las características predictivas y la variable objetivo. Para este caso se utilizó el regularizador más comúnmente utilizado, utilizando una función cuadrática de error y un alpha con valor a 1. Para esta selección de modelo fue importante entender la herramienta computacional que proporciona RidgeRegression en sci-kit learn y se utilizaron los valores predeterminados de esta librería, incluido un solucionador automático que se ajusta a este tipo de datos.

El tercer modelo utilizado para esta práctica se conoce como Support Vector Machines, el cual funciona correlacionando datos a un espacio de características de grandes dimensiones, de forma que los puntos de datos no se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Por lo que se detecta un separador entre características para predecir el grupo al cual pertenece el nuevo registro. Para estos datos, se encontró el mejor modelo que separa dichas características de forma empírica entre varias formas de un hiperplano como lo son el lineal, el polinómico, la función de base radial y un kernel sigmoide.

Luego de haber encontrado el mejor modelo para cada uno, se hizo una comparación entre los 3 modelos basado en 4 indicadores: la precisión, que indica la proporción de valores que fueron correctamente clasificados, el recall o recuperación, que indica la proporción de valores que son de la clase que son identificados de forma correcta, la exactitud o accuracy, que indica la cantidad de predicciones correctas con respecto al número total.

Luego de correr los modelos, se anotaron todos los indicadores en la siguiente tabla, y en la sección de resultados se dará el veredicto sobre cuál fue el mejor modelo en términos de su rendimiento en métricas de calidad

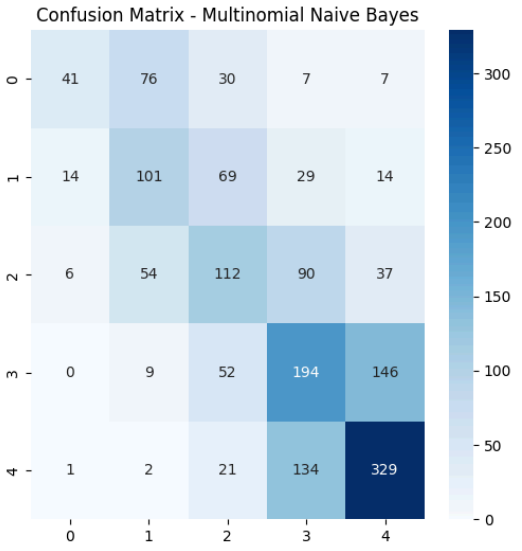
| | Naive Bayes | Clasificador Ridge | Support Vector Machines |
|---|---|---|---|
| Exactitud | 0,49 | 0,42 | 0,43 |
| Precisión | 0,5 | 0,4 | 0,41 |
| Recall | 0,45 | 0,37 | 0,35 |
| F1-Score | 0,46 | 0,38 | 0,34 |

## 5. Resultados.

El mejor modelo para implementar es Naive Bayes Multinomial. Sus resultados son los siguientes:

- Precisión: 0.5035
- Exactitud (Accuracy): 0.4933
- Recuperación (Recall): 0.4467
- Puntuación F1 (F1 Score): 0.4563

En la siguiente matriz de confusión se permite observar qué valores fueron bien clasificados y mal clasificados con respecto a la clase a la que corresponde dicha frase teóricamente




Naive Bayes es una opción sólida para la clasificación, especialmente si se tienen en cuenta las limitaciones y se aplican pre procesamientos adecuados a los datos.

Escoger este modelo puede ayudar dentro del modelo a.

- Esta puede clasificar si los comentarios son positivos, negativos o neutrales. Esto ayudaría a identificar áreas específicas que requieren mejoras o inversiones.
- Agrupar a los clientes satisfechos en un grupo y a los insatisfechos en otro. Luego, la empresa podría dirigir sus inversiones hacia las áreas que afectan más a los grupos insatisfechos.
- Predecir preferencias, esto ayudaría a adaptar las ofertas y servicios de la empresa en función de esas preferencias.

## 6. Mapa de actores

| Rol dentro de la empresa | Tipo de actor | Beneficio | Riesgo |
|---|---|---|---|
| Ministerio de Comercio, Industria y Turismo | Patrocinador | Mejora en la promoción y gestión de destinos turísticos | Inversión en proyectos no rentables o ineficaces |
| COTELCO y cadenas hoteleras | Beneficiarios | Incremento en la ocupación y satisfacción del cliente | Daño a la reputación por sitios mal valorados |
| Hoteles pequeños | Beneficiarios | Oportunidades de mejora y competitividad | Riesgo de inversión en mejoras no efectivas |
| Científicos de datos e ingenieros | Ejecutores | Desarrollo profesional y contribución al sector | Presión por resultados y precisión en el análisis |

## 7. Trabajo en equipo

Resumen Ejecutivo: Este documento resume las actividades y responsabilidades del equipo encargado del proyecto de análisis turístico. El objetivo es evaluar las características de los sitios turísticos y su impacto en la popularidad y recomendaciones de los turistas.

**Roles y Responsabilidades:**

- Líder del Proyecto (Felipe Rueda y Santiago Pardo): Responsable de la gestión general del proyecto, incluyendo la planificación de reuniones, distribución equitativa de tareas y la subida de entregables. Tiene la autoridad final en decisiones sin consenso.
- Líder de Negocio (Luis Felipe Plazas y Felipe Rueda): Asegura que el proyecto resuelva el problema identificado y esté alineado con la estrategia del negocio. Coordina con expertos estadísticos y comunica efectivamente el producto.
- Líder de Datos (Santiago Pardo): Gestiona los datos utilizados en el proyecto y asigna tareas relacionadas con los mismos, asegurando su disponibilidad para el equipo.
- Líder de Analítica (Santiago Pardo): Supervisa las tareas analíticas y verifica que los entregables cumplan con los estándares de análisis y restricciones del proyecto.

**Cronograma de Reuniones:**

- Reunión de Lanzamiento y Planeación: Realizada el 19 de marzo para definir roles y metodologías de trabajo, además de generar ideas para la resolución del proyecto.
- Reunión de Ideación: El 3 de abril, se discutió la organización y los roles beneficiados por la solución analítica tras explorar los datos.
- Reuniones de Seguimiento: Se llevó a cabo una reunión semanal el 5 de abril para monitorear el progreso y ajustar estrategias según sea necesario.
- Reunión de Finalización: El 6 de abril, se revisaron los resultados finales y se discutieron los pasos a seguir.
- Herramientas Utilizadas: Para la gestión de tareas se utilizó Whatsapp y Zoom para una mejor comunicación, permitiendo un seguimiento claro y organizado del progreso del proyecto.

**Repartición de los 100 puntos:**

- Felipe Rueda : 34 puntos
- Santiago Pardo: 33 puntos
- Luis Felipe Plazas: 33 puntos

Como aspecto a mejorar, en próximas entregas se hará una implementación un código de más calidad y se hará una mejor comunicación para realizar el trabajo conjunto, tanto con los expertos en datos como los analistas de datos.

Conclusiones: El equipo ha demostrado una colaboración efectiva, cumpliendo con los roles asignados y manteniendo una comunicación constante. Los resultados obtenidos serán fundamentales para mejorar la experiencia turística y fomentar el turismo en Colombia.

## 8. Referencias


- https://theblackboxlab.com/2022/03/30/modelos-naive-bayes/#:~:text=Naive%20Bayes%20es%20uno%20de,cuando%20hablamos%20de%20predicciones%20multiclase.
- https://interactivechaos.com/es/manual/tutorial-de-machine-learning/regresion-ridge
- https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works
- https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall?hl=es-419
- https://developers.google.com/machine-learning/crash-course/classification/accuracy?hl=es-419.