Proyecto 1 – Etapa 1: Analítica de Textos

Universidad de los Andes Departamento de Ingeniería de Sistemas y Computación ISIS 3301 Inteligencia de Negocios 202320

Contenido

5
6
8
con el
11

1. Entendimiento del negocio y enfoque analítico.

Oportunidad de Negocio

El proyecto de análisis de sitios turísticos en Colombia, liderado por el Ministerio de Comercio, Industria y Turismo en colaboración con COTELCO y cadenas hoteleras como Hilton y Holiday Inn, busca comprender las características que hacen atractivos a los destinos turísticos para visitantes locales y extranjeros. Este análisis permitirá identificar elementos clave que influyen en la experiencia del turista, desde la accesibilidad hasta la calidad de los servicios ofrecidos. Además, se compararán sitios turísticos populares con aquellos menos recomendados para identificar áreas de mejora y diseñar estrategias para mejorar la experiencia del turista. Un objetivo crucial del proyecto es desarrollar un sistema de calificación basado en las opiniones de los turistas para evaluar la satisfacción del visitante y proporcionar datos para la toma de decisiones orientadas a la mejora continua de los destinos turísticos. El éxito en la implementación de este sistema permitirá identificar oportunidades de mejora y adaptar estrategias para satisfacer las necesidades cambiantes de los turistas, beneficiando así al sector turístico colombiano en general. Además, el Ministerio de Turismo emerge como un actor clave beneficiado por el proyecto, ya que un aumento en la popularidad y calidad de los destinos turísticos puede tener efectos positivos en el desarrollo económico y social del país. Estimular el turismo y mejorar la experiencia del visitante puede generar nuevas oportunidades de empleo, impulsar sectores relacionados como la hostelería y la restauración, y fortalecer la reputación de Colombia como un destino turístico atractivo a nivel nacional e internacional. Este proyecto respalda los esfuerzos del gobierno para promover el

turismo y mejorar la imagen del país en el exterior, demostrando
así los grandes beneficios potenciales para el gobierno y la
sociedad en su conjunto.

Enfoque Analítico

Considerando el contexto proporcionado por el negocio y los objetivos delineados para el proyecto, se propone un enfoque analítico predictivo basado en el aprendizaje supervisado. Este enfoque permitirá extraer insights significativos al relacionar los testimonios de los usuarios con las calificaciones asociadas a los destinos turísticos. Al adoptar una tarea de clasificación sobre los datos recopilados, se buscará identificar patrones y tendencias en las reseñas que puedan prever la calificación asignada por los turistas. Para lograr esto, se plantea la implementación de tres técnicas diferentes: RandomForest, LogisticRegression y MultinomialNB.

La elección de estas metodologías se basa en su capacidad para abordar distintos aspectos del problema y su flexibilidad para adaptarse a diferentes tipos de datos y relaciones. RandomForest destaca por su capacidad para manejar conjuntos de datos complejos y su resistencia al sobreajuste, lo que lo convierte en una opción sólida para este proyecto. LogisticRegression, por otro lado, es conocido por su eficacia en problemas de clasificación binaria o multiclase, y podría ser útil para modelar la relación entre las características de las reseñas y las calificaciones asignadas. Por último, MultinomialNB es especialmente adecuado para datos de texto y características discretas, lo que lo convierte en una opción prometedora dadas las características de las reseñas de los sitios turísticos.

La adopción de estas tres técnicas proporcionará una perspectiva holística y completa sobre el problema, permitiendo la construcción de modelos variados y la evaluación de su desempeño a través de diversas métricas de calidad. Esta aproximación ayudará a identificar el modelo más óptimo y efectivo para predecir las calificaciones de las reseñas, lo que contribuirá significativamente a mejorar la experiencia turística en Colombia.

Organización y rol que se beneficia

Teniendo en cuenta que todos los actores implicados están relacionados entre sí, podemos afirmar lo siguiente y es que todas las organizaciones implicadas en el proyecto se verán beneficiadas de las siguientes maneras. El Ministerio fortalecerá su papel como promotor del turismo al mejorar la calidad de los destinos turísticos, mientras que COTELCO representará los intereses del sector hotelero, los cuales tendrán como insumo importante las opiniones de los turistas. Las cadenas hoteleras y los pequeños hoteles obtendrán información valiosa para adaptar sus servicios y estrategias a las necesidades del mercado, lo que puede resultar en un aumento de la demanda y una mayor competitividad en el sector.

Al participar en el proyecto, estas organizaciones pueden colaborar para mejorar la experiencia del turista en Colombia. Se espera que esto tenga un impacto positivo en el sector turístico, aumentando la demanda de alojamiento y mejorando la ocupación hotelera. Además, al implementar mejoras sugeridas en los destinos turísticos locales, los pequeños hoteles pueden fortalecer su competitividad y aumentar su visibilidad en el mercado. Teniendo en cuenta lo anterior se identifican los beneficios en lo siguientes roles:

Transparencia en la toma de decisiones: El Ministerio de Comercio, Industria y Turismo de Colombia se beneficiará al mejorar la transparencia en la toma de decisiones relacionadas con el desarrollo y promoción turística del país. Al tener acceso a datos y análisis detallados sobre las características y calificaciones de los sitios turísticos, el Ministerio podrá fundamentar sus decisiones en información objetiva y precisa, lo que aumentará la confianza tanto de los ciudadanos como de los inversores en las políticas y estrategias implementadas.

Eficiencia en la gestión de recursos: Tanto la Asociación Hotelera y Turística de Colombia (COTELCO) como las cadenas hoteleras y los pequeños hoteles se beneficiarán al mejorar la eficiencia en la gestión de recursos. Al tener una comprensión más clara de las preferencias y necesidades de los turistas, estas organizaciones podrán dirigir sus inversiones y esfuerzos hacia áreas específicas que mejoren la experiencia del visitante, lo que resultará en una mejor utilización de los recursos disponibles y un mayor retorno de la inversión.

Mejora en la calidad de los servicios: Todos los actores involucrados se beneficiarán de una mejora en la calidad de los servicios ofrecidos en los destinos turísticos. Al identificar áreas de mejora y aplicar estrategias para aumentar la satisfacción del turista, se espera que los servicios de alojamiento, transporte, gastronomía y actividades turísticas en general alcancen niveles más altos de calidad y excelencia, lo que beneficiará tanto a los visitantes como a las empresas y trabajadores del sector. En resumen, el proyecto de análisis de sitios turísticos beneficia a diferentes actores del sector turístico colombiano al colaborar para mejorar la calidad y competitividad de los destinos turísticos del país. Esta colaboración promueve un desarrollo sostenible y fortalece la reputación de Colombia como un destino turístico seguro y atractivo tanto a nivel nacional como internacional.

Contacto con experto externo al provecto Nuestro contacto experto para el proyecto es Manuela Rodriguez Vargas.

Iniciamos el contacto a través de su correo electrónico: m.rodriguezv2@uniandes.edu.co

Gracias a esto, intercambiamos números de teléfono y acordamos una reunión para explicarle en detalle el proyecto que se va a realizar e iniciar el trabajo conjunto. Por lo tanto, los canales a utilizar serán WhatsApp y zoom para el caso de la reunión virtual. Fecha de la reunión: 5 de Abril Se planea otra reunión para la realización y consecución de los resultados necesarios para la etapa 2.

2. Entendimiento y preparación de los datos.

Perfilamiento: Durante el proceso de perfilamiento de los datos, hemos explorado un conjunto de registros compuesto por 7,875 reseñas proporcionadas por diversos individuos, todas relacionadas con diferentes lugares turísticos visitados. Cada entrada de datos consta de dos columnas distintas: una denominada "Review", que contiene el texto de la reseña expresada por el individuo, y otra titulada "Class", que representa la calificación asignada por el turista al lugar visitado mediante un valor numérico. En este contexto, nuestra principal columna de interés es la primera, ya que nos brinda información fundamental para establecer la relación entre las reseñas y las calificaciones correspondientes. A través del análisis de estas reseñas, buscamos desarrollar un modelo predictivo capaz de anticipar las calificaciones asociadas a una determinada reseña.

Análisis de calidad de datos:

- **Completitud:** Durante el análisis de completitud, se observó el porcentaje de valores nulos en cada una de las columnas de los datos. Se determinó que ambas columnas están completas, con un 0% de incompletitud. Esto indica que los 7875 registros están completos en su totalidad.
- Unicidad: En el análisis de unicidad, se buscó la presencia de registros duplicados, encontrando un total de 73 duplicados. Es importante destacar que este aspecto es relevante, especialmente para nuestra primera columna de texto, donde los valores pueden repetirse en diferentes registros debido a la naturaleza de las reseñas. Además, en el caso de la columna numérica, donde los valores solo pueden ser 1 o 5, se observaron repeticiones en los diferentes registros.
- **Consistencia:** Se identificaron 3 registros en la columna "Review" que estaban en inglés. Además, se encontró un error en la librería utilizada para el análisis del idioma, que detectó algunos registros como pertenecientes a un idioma distinto al español, pero posteriormente se confirmó que eran en español.
- Validez: Para el análisis de validez, se consideraron todos los registros como válidos, dado que no se dispone de un diccionario de datos para realizar comparaciones. En el caso de la columna numérica, se verificó que todos los valores estuvieran dentro del rango esperado, es decir, entre 1 y 5, conforme a las calificaciones posibles para un lugar turístico.

Preparación de datos:

Para la implementación de los modelos que nos ayudarán a abordar la tarea propuesta, es esencial realizar una preparación exhaustiva de los datos. En este proceso, se detectaron problemas previos que requerían corrección, incluyendo textos que no estaban en español y registros duplicados. Por tanto, se procedió a eliminar estos cuatro registros y los duplicados respectivos, garantizando así que los datos estén libres de impedimentos que puedan afectar la construcción de los modelos y la calidad de los resultados.

Posteriormente, se realizó una serie de transformaciones en los datos para asegurar su adecuación para su posterior procesamiento. Esto implicó realizar una limpieza rigurosa de los textos, que incluyó la sustitución de números por su representación textual, la eliminación de caracteres no pertenecientes al alfabeto latino, así como la eliminación de espacios y puntuación innecesarios. Una vez completada esta etapa, se aplicó la lematización, un proceso que reduce todas las palabras a su forma base o raíz, generando así un conjunto de tokens representativos de cada mensaje. Finalmente, se procedió a dividir los datos en conjuntos de entrenamiento y prueba en una proporción de 80-20, asegurando que el modelo tenga suficientes datos para aprender y evaluar su desempeño de manera efectiva. Luego, se utilizó la técnica TFidf para la vectorización de los datos, generando vectores numéricos que representan la frecuencia relativa de cada palabra en relación con el corpus completo de mensajes. Estos vectores, junto con sus correspondientes calificaciones, serán utilizados en la construcción y evaluación de los modelos de aprendizaje automático.

3. Modelado y Evaluacion

Para la realización y el alcance de los objetivos planteados para este proyecto, se desarrollaron tres diferentes modelos en base a tres diferentes algoritmos de clasificación apropiados para el contexto y los datos que se prepararon para esto, a continuación, explicaremos cada uno de ellos:

• Algoritmo Logistic Regression (Juan Camilo Pacheco): La regresión logística es un método estadístico utilizado para predecir una variable categórica binaria (como sí/no, éxito/fracaso) en función de una o más variables predictoras. Utiliza la función logística para modelar la relación entre las variables predictoras y la probabilidad de pertenecer a una categoría específica. Para la realización de este algoritmo se deben determinar 3 hiperparámetros: el tipo de regularización (penalty), la fuerza de la regularización (C) y el algoritmo de optimización que se utilizará (solver). Mediante el uso de un parameter grid, se encontraron los hiperparámetros óptimos: penalty: Ridge, C: 10000000000, solver: newton-cg. Con estos parámetros, se entrenó el modelo y se procedió a probarlo tanto en el set de entrenamiento como en el de prueba. La matriz de confusión y el reporte de clasificación encontrados son los siguientes:

[[55	54	20	8	9]
[28	80	74	28	28]
					41]
[3	13	54	171	160]
[0	2	24	92	346]]

	precision	recall	f1-score	support
1	0.57	0.38	0.45	146
2	0.39	0.34	0.36	238
3	0.43	0.40	0.41	325
4	0.44	0.43	0.44	401
5	0.59	0.75	0.66	464
accuracy			0.50	1574
macro avg	0.48	0.46	0.46	1574
weighted avg	0.49	0.50	0.49	1574

Como podemos ver, se obtuvo un accuracy de 0.50, indicando que el modelo no es muy bueno para predecir correctamente las calificaciones de las reseñas. De igual manera se obtiene un recall de aproximadamente 0.5 indicando que el modelo solo logra identificar correctamente el 50% de todas las instancias. Sin embargo más adelante veremos que con este modelo obtivimos el mejor resultado.

• Algoritmo MultinomialNB (Santiago Rodriguez Bernal):

El algoritmo MultinomialNB (Naive Bayes Multinomial) es un clasificador de aprendizaje automático ampliamente utilizado en problemas de clasificación de texto. Este método se basa en el teorema de Bayes y opera bajo el supuesto de independencia entre las características. Se emplea para estimar la probabilidad de que un dato pertenezca a una de varias clases posibles, teniendo en cuenta la frecuencia de ocurrencia de características, como palabras en un documento. Es particularmente útil en aplicaciones de procesamiento de lenguaje natural, donde las características se refieren a términos o palabras y se evalúa cuántas veces aparecen en un texto. A través de la implementación de este algoritmo, se obtuvieron las siguientes métricas en nuestro modelo:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	146
2	0.30	0.11	0.17	238
3	0.31	0.19	0.24	325
4	0.29	0.30	0.30	401
5	0.46	0.87	0.61	464
accuracy			0.39	1574
macro avg	0.27	0.30	0.26	1574
weighted avg	0.32	0.39	0.33	1574

Como se puede observar se obtuvo un accuracy y un recall de tan solo 0.39 lo que implica que el modelo no es bueno para la tarea propuesta ya que no logra identificar correctamente la mayoría de las reseñas analizadas.

Algoritmo Random Forest (Juan Camilo Pacheco y Santiago Rodriguez Bernal): Random Forest es un algoritmo usado para tareas de clasificación y regresión. Este algoritmo es basado en árboles de decisión el cual logra combinar la salida de varios árboles de decisión para poder lograr a un resultado. El funcionamiento de este algoritmo se basa en tres hiperpárametros que son el tamaño del nodo, la cantidad de árboles y la cantidad de características muestreadas. Para su implementación, primero, importar las respectivas bibliotecas (en este caso RandomForestClassifier), luego se realizó la creación del modelo en donde el random state se especifica en 0 para afianzar la reproducibilidad. Una vez con esto, se entrenan el modelo, para así poder comenzar a realizar las particiones para la validación cruzada, en donde se usó 5 particiones teniendo en cuenta la cantidad de datos que se está manejando. Por otro lado, se realiza la definición de los hiperpárametros para realizar una búsqueda sistemática de los mejores hiperpárametros para este modelo y con este resultado, lograr arrojar el informe de clasificación con sus respectivos resultados. Por último, pudimos obtener los siguientes datos:

	· 1							
Classification Report								
	precision	recall	f1-score	support				
1	0.50	0.17	0.26	146				
2	0.36	0.35	0.36	238				
3	0.38	0.27	0.32	325				
4	0.41	0.32	0.36	401				
5	0.49	0.79	0.60	464				
accuracy			0.44	1574				
macro avg	0.43	0.38	0.38	1574				
weighted avg	0.43	0.44	0.41	1574				

En este caso obtuvimos un recall y accuracy similar al obtenido con el algoritmo Logistic Regression, sin superarlo, indicando que el modelo no es muy bueno para realizar la tarea propuesta.

Al analizar los 3 modelos implementados, ninguno de ellos logro identificar correctamente más del 50% de las reseñas. Sin embargo, el modelo de Logistic Regression tuvo mayor precisión en sus resultados.

4. Resultados.

Tras la evaluación de los modelos implementados, se determinó que el modelo de regresión logística fue el más adecuado para predecir las calificaciones de las reseñas de los sitios turísticos, obteniendo un accuracy de 0.496, el más alto en comparación con los otros modelos evaluados. Este resultado indica que el modelo de regresión logística tiene una mayor capacidad para capturar las relaciones entre las características de las reseñas y las calificaciones asignadas por los turistas.

En términos más cualitativos, la eficacia del modelo de regresión logística sugiere la posibilidad de automatizar el proceso de evaluación de las reseñas de los turistas, lo que facilitaría la identificación de las características clave que influyen en la experiencia

del visitante. Al entender mejor las razones detrás de las calificaciones de las reseñas, se pueden diseñar estrategias específicas para mejorar la experiencia del turista en los destinos turísticos de Colombia.

Aunque el modelo de regresión logística demostró ser el más preciso en la predicción de las calificaciones de las reseñas, se identificaron limitaciones en su capacidad para predecir con precisión las calificaciones intermedias. Esto sugiere que el modelo puede beneficiarse de la incorporación de características adicionales o el uso de técnicas más avanzadas de procesamiento de lenguaje natural para capturar la complejidad de las opiniones de los turistas.

Por otro lado, se ha determinado que el modelo de regresión logística fue el más efectivo entre los evaluados, aunque su precisión, medida por el F1-score, reveló una calidad insuficiente con un valor de 0.50. Este bajo rendimiento se atribuye al desequilibrio en la distribución de los datos de entrenamiento, donde la abundancia de registros con una calificación de 5 en comparación con las calificaciones más bajas afecta la capacidad predictiva del modelo. Para mejorar su calidad, se recomienda abordar este desequilibrio mediante técnicas de re-muestreo y explorar otros enfoques de modelado, como modelos de ensamble o redes neuronales.

Además, se sugiere realizar un análisis de sentimiento sobre las reseñas de los usuarios para complementar la clasificación de las calificaciones de los sitios turísticos. Este análisis permitiría identificar el tono emocional de las opiniones expresadas en las reseñas, proporcionando una visión más completa de la percepción de los turistas hacia los destinos turísticos. Asimismo, el análisis de sentimiento podría ayudar a identificar tendencias y patrones en las opiniones de los usuarios a lo largo del tiempo, lo que sería útil para detectar áreas de mejora y diseñar estrategias para mejorar la experiencia turística en los destinos.

En conclusion, la combinación de la clasificación de las calificaciones de los sitios turísticos y el análisis de sentimiento sobre las reseñas de los usuarios proporcionaría una comprensión más completa de la percepción de los turistas y permitiría una toma de decisiones más informada en el sector turístico. Estas estrategias permitirían abordar los desafíos de precisión del modelo y mejorar la calidad de la información disponible para los responsables de la toma de decisiones en el ámbito turístico.

 Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido

Organización/ Apartado	Rol dentro de la empresa	Tipo de Actor	Beneficio	Riesgo
Equipo de Investigación de la Universidad de los Andes	Investigadore s	Usuarios Internos	Los investigadores de la Universidad de los Andes emplean el producto de datos para realizar investigaciones académicas centradas en el sector turístico en Colombia. Este enfoque contribuye al avance del conocimiento en el campo del turismo, fortaleciendo la comprensión de los factores que influyen en la industria turística del país y sus implicaciones para el desarrollo económico y social.	Si los datos no son precisos o completos, las investigaciones académicas basadas en ellos podrían carecer de fiabilidad.
Estudiantes de la Universidad de los Andes	Participantes	Usuarios Internos	Los estudiantes pueden aprender sobre los retos y posibles mejoras que tiene el turismo en el país a través de proyectos y análisis basados en el producto de datos, lo que enriquece su educación y genera conciencia sobre las oportunidades dentro del sector	Si la participación de los estudiantes en la recopilación o análisis de datos no se gestiona adecuadamente, podría haber desafíos logísticos o éticos.
Ministerio de Turismo de Colombia	Departament o de Planificación Turística	Cliente	El Ministerio de Turismo podría utilizar los resultados del modelo analítico para tomar decisiones informadas sobre la promoción de destinos turísticos, la asignación de recursos y la elaboración de políticas para impulsar el turismo en Colombia. Esto podría resultar en un aumento del flujo turístico, la generación	Existe el riesgo de que el Ministerio de Turismo no tenga la capacidad suficiente para interpretar y aplicar adecuadamente los resultados del modelo analítico. Esto podría llevar a decisiones inadecuadas o malentendidos en la planificación turística, lo que podría afectar

			de ingresos y el desarrollo económico en las regiones turísticas del país.	negativamente la reputación del Ministerio y el desarrollo del sector turístico en Colombia.
Asociación Hotelera y Turística de Colombia (COTELCO)	Departament o de Desarrollo de Productos Turísticos	Proveedor	COTELCO podría utilizar los resultados del modelo analítico para identificar tendencias en las preferencias de los turistas y desarrollar ofertas turísticas más atractivas y adaptadas a las necesidades del mercado. Esto podría resultar en un aumento de la ocupación hotelera y la rentabilidad para los establecimientos afiliados a la asociación.	Existe el riesgo de que los proveedores no estén dispuestos o no tengan la capacidad de implementar los cambios sugeridos por el modelo analítico. Esto podría limitar el impacto de las mejoras propuestas y llevar a una pérdida de competitividad en el mercado turístico.
Agencia Nacional de Promoción Turística de Colombia (FONTUR)	Departament o de Marketing Turístico	Cliente	FONTUR podría utilizar los resultados del modelo analítico para optimizar sus estrategias de promoción turística, enfocando los recursos en los destinos con mayor potencial de atracción de turistas. Esto podría aumentar la eficacia de las campañas de marketing y contribuir a un mayor posicionamiento de Colombia como destino turístico.	Existe el riesgo de que la implementación de estrategias basadas en los resultados del modelo no genere el impacto esperado, lo que podría llevar a una subutilización de los recursos invertidos en promoción turística y a una disminución de la visibilidad de Colombia como destino turístico.

Empresas de Transporte Turístico en Colombia	Departament o de Operaciones y Logística	Proveedor	Las empresas de transporte turístico podrían utilizar los resultados del modelo analítico para planificar rutas y horarios más eficientes, optimizando la oferta de transporte hacia los destinos turísticos más populares. Esto podría mejorar la experiencia de viaje de los turistas y aumentar la satisfacción del cliente.	Existe el riesgo de que las empresas de transporte no cuenten con la infraestructura necesaria para adaptarse a las demandas generadas por las recomendaciones del modelo analítico. Esto podría resultar en retrasos, cancelaciones o deficiencias en el servicio, lo que afectaría negativamente la experiencia de los turistas y la reputación de las empresas.
Agencias de Viajes en Colombia	Departament o de Ventas y Marketing		Las agencias de viajes podrían utilizar los resultados del modelo analítico para personalizar sus ofertas y recomendar destinos y actividades turísticas más relevantes para cada tipo de cliente. Esto podría mejorar la experiencia de reserva de los turistas y aumentar la fidelidad del cliente hacia la agencia.	Existe el riesgo de que las agencias de viajes no cuenten con la capacitación adecuada para interpretar y aplicar los resultados del modelo analítico de manera efectiva. Esto podría llevar a recomendaciones erróneas o malentendidos en la comunicación con los clientes, lo que podría afectar la satisfacción del cliente y la reputación de la agencia.

6. Trabajo en equipo

Estudiantes	Rol	Horas trabajadas	Tareas hechas	Desafíos	Puntaje
Santiago Rodriguez	Líder de	14	Algoritmo –	Investigación para	50
-	proyecto		Documento -	la preparación de	
			Vídeo	datos	
Juan Camilo Pacheco	Líder de	14	Algoritmo	Investigación de	50
	negocio		Documento –	los algoritmos y	
			Vídeo	mejores modelos	

Después de llevar a cabo una reunión de equipo, se decidió asignar un puntaje de 50 a cada miembro, con el fin de lograr una suma total de 100. Esta distribución equitativa se basó en el principio de dividir el trabajo de manera equitativa para aumentar la eficiencia en la realización del proyecto. En cuanto al tiempo dedicado al proyecto, se determinó que fue de 14 horas en total, ya que cada miembro trabajó dos horas diarias durante un período de 7 días.

Durante la discusión sobre los aspectos a mejorar para la próxima entrega, se identificó como punto crucial la necesidad de comenzar el proyecto con más anticipación para poder obtener mejores resultados. Se reconoció que el buen trabajo en equipo y la comunicación efectiva fueron aspectos destacados en esta entrega y se acordó que deben mantenerse como prioridades para futuros proyectos. Es importante destacar que se llevó a cabo una reunión inicial para compartir esta propuesta y definir los procesos de trabajo, especialmente en lo que respecta a la comunicación para garantizar la ejecución efectiva del proyecto. Durante la reunión final, se concretaron los últimos detalles de cada uno de los entregables, asegurando así su calidad y completitud.