Universidad de los Andes

Proyecto 1 Etapa 1

Inteligencia de Negocios

GRUPO PROYECTO 3
Maria Alejandra Estrada García - 202021060
Marilyn Stephany Joven Fonseca - 202021346
Santiago Martínez Novoa - 202112020

Contents

1 Entendimiento del negocio y enfoque analítico

En Colombia, el turismo es uno de los sectores más dinámicos de la economía después del petróleo y el carbón, representando alrededor del 6,1% del PIB de acuerdo con estudios realizados por MasterCard y el Centro de Desarrollo Internacional de la Universidad de Harvard (CID) en el 2016. Este sector viene creciendo con tasas superiores a la mayoría de los demás paises de la región, convirtiéndose en un importante factor para el desarrollo de la economía y de las regiones.

Actualmente, a causa del alto avance tecnológico y la gran influencia del internet en la vida de las personas, las reseñas para un hotel se ha convertido en una pieza clave.Esto se debe a que el 95% de los que viajan por placer leen las reseñas antes de reservar sus vacaciones y el 49% no se arriesgarían nunca a reservar un hotel que no tenga "opiniones". Esto es una clara muestra de que las reseñas tienen un impacto directo en los ingresos de un hotel.

Teniendo en cuenta este conteto, los **objetivos** de este proyecto son los siguientes:

- Identificar las características que hacen atractivo un hotel sobre el resto y como estas afectan la opinión y la probabilidad de recomendación de los usuarios.
- Diseñar un modelo de aprendizaje automático que, dado una reseña de un usuario, permita a las cadenas hotelera y a COTELCO clasificarlo como relacionado a una calificación del hotel en una escala entera de 1 a 5.

Se establecen dos **criterios de éxito** para este proyecto:

1. Que el modelo escogido sea capaz de recibir una reseña nueva y que pueda clasificarla en alguna de las 5 clases.
2. Que el modelo exhiba una probabilidad mayor al 50% de que la clasificación sea correcta.

Oportunidad/problema Negocio	Mejorar la promoción y gestión turística en Colombia mediante el análisis de las características de los sitios turísticos para identificar qué los hace atractivos o poco recomendados. Esto permitirá tomar decisiones informadas para aumentar la popularidad de los destinos y fomentar el turismo en el país.

Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Para este proyecto de procesamiento de lenguaje natural (NLP), nos centraremos en el enfoque basado en inteligencia artificial (IA) para analizar las características de los sitios turísticos y comparar su atractivo para los turistas. El objetivo principal del negocio es identificar oportunidades de mejora en el turismo mediante técnicas de aprendizaje automático. Comenzaremos con un análisis de sentimientos utilizando Clasificación de Texto con SVM para determinar el sentimiento asociado con las reseñas de los sitios turísticos. Luego, emplearemos el algoritmo de Random Forest para identificar las características más relevantes que inciden en la satisfacción del turista y la popularidad de los destinos turísticos. Este método se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y combina sus predicciones para obtener una predicción más precisa y robusta. Finalmente, utilizaremos la Regresión Logística para predecir la probabilidad de que un sitio turístico reciba una calificación positiva, neutral o negativa en función de sus características específicas.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización que se beneficiaría con esta oportunidad es el Ministerio de Comercio, Industria y Turismo de Colombia. Su rol dentro de esta oportunidad sería el de liderar y coordinar las acciones destinadas a mejorar la promoción turística en el país. Además, sería responsable de utilizar los resultados del análisis de datos para diseñar políticas y estrategias que impulsen el turismo nacional e internacional, en colaboración con otras entidades gubernamentales y actores del sector turístico.

Contacto con experto externo al proyecto y detalles de la planeación	Se tuvo contacto con 1 estudiante de estadística, Juan Francisco Bernal (j.bernalc@uniandes.edu.co), con quien se tendrán reuniones constantes de manera presencial a partir de la primera reunión el sábado 6 de abril a las 5:00pm horas, en la cual se le informará del proceso del proyecto y de los resultados de esta primera parte, y se tendrá una comunicación fluida usando como canal el grupo de WhatsApp

1.1 Enfoque analítico

El procesamiento de lenguaje natural (NLP) funciona mediante diversos enfoques que abarcan desde algoritmos basados en inteligencia artificial (IA) hasta métodos estadísticos y reglas predefinidas. Para este proyecto, nos centraremos en el enfoque basado en IA, que implica el uso de algoritmos de aprendizaje automático para procesar y comprender el lenguaje humano.

El objetivo principal del negocio es analizar las características de los sitios turísticos que los hacen atractivos para los turistas y compararlos con aquellos que han obtenido bajas recomendaciones, con el fin de identificar oportunidades de mejora y fomentar el turismo.

En este caso se realizarán 1 técnicas principales de aprendizaje automático las cuales pertenecen a un aprendizaje supervisado que es un tipo de aprendizaje donde se entrena al modelo utilizando un conjunto de datos que incluyen ejemplos de entrada y la salida deseada correspondiente y datos etiquetados. En esta categoría se emplea clasificación que se tratan respectivamente de predecir valores numéricos continuos y de asignar categorías a las instancias de datos.Es decir, la clasificación es otra técnica de aprendizaje supervisado que se utiliza para asignar categorías a las instancias de datos. En este caso, el modelo aprende a clasificar los datos en diferentes clases o categorías basándose en las características de entrada.

1.2 Estudiantes curso estadística

- Juan Francisco Bernal (j.bernalc@uniandes.edu.co, 202010643)

2 Entendimiento y preparación de los datos

Se analizó la información básica del conjunto de datos, que consta de 7875 opiniones en dos categorías principales. La primera, llamada "Review", contiene todas las experiencias de los clientes durante sus viajes, mientras que la segunda, "Class", asigna calificaciones del 1 al 5 para resumir la calidad general de la experiencia. Notamos que algunas palabras están escritas de forma incorrecta o informal, como "pq" y "q" en lugar de "porque" y "que", respectivamente. También hay errores con las tildes en palabras como "está".

Al revisar los datos más detenidamente, descubrimos que no hay información faltante, pero encontramos 71 entradas duplicadas. Para mantener solo una copia de cada una, planeamos eliminar las duplicadas y conservar solo la primera.

Al observar las calificaciones en la columna "Class", notamos que la mayoría de las opiniones tienen una puntuación de 5, lo que indica que la mayoría son muy positivas. A medida que la calificación disminuye, la cantidad de opiniones también disminuye, lo que sugiere que hay menos opiniones negativas en comparación con las positivas.

También revisamos las reseñas en la columna "Reviews" y notamos que algunas son muy largas, lo que puede dificultar su análisis. Decidimos limitarlas a un máximo de 500 palabras para simplificar. Después de este filtro, quedaron 5897 reseñas, con un promedio de 247.15 palabras cada una.

Además, quisimos identificar qué idiomas se usaron en las reseñas. Descubrimos que el 99% están en español, con solo una pequeña cantidad en otros idiomas como inglés. Debido a su escasez, decidimos eliminar estas reseñas en otros idiomas para centrarnos en las que están en español.

2.1 Decisiones

Teniendo en cuenta la preparación y el entendimiento se tomaron las siguientes decisiones.

1. **Eliminación de duplicados:** Teniendo en cuenta la cantidad de duplicados se eliminaron los duplicados excepto por el primero de estos para aun así tener uno de los registros.
2. **Identificación del idioma:** Se procede a examinar los diferentes idiomas utilizados y se eliminan las reseñas con algun idioma diferente al español.
3. **Limpieza de caracteres:** Como los textos pueden tener simbolos como puntos, comas, y otros caracteres, se eliminan estos para obtener un texto de solo letras, ignorando mayusculas o tildes.
4. **Eliminar palabras sin significado:** En los textos se presentan muchas palabras que pueden no aportar sentido al análisis, y se toma la decisión de eliminar cualquier palabra considerada como una palabra vacía (stop word).
5. **Verificación vocabulario y conteo de ocurrencias:** Para obtener una visión general del vocabulario se calcula la ocurrencia de cada palabra. Por último se eliminaron todas aquellas palabras con una ocurrencua de 1 en todo el dataset para no tomar en cuenta palabras raras pues, en caso de ser utilizadas solo una vez no son útiles para el análisis.
6. **Normalización:** Se realiza Stemming y Lematización para reducir palabras a su raíz de manera que puedan ser interpretadas en más de una de sus formas.
7. **Vectorización de texto:** Teniendo en cuenta que las redes neuronales y las máquinas de vectores de soporte (SVM), requieren que las entradas sean en forma numérica se realiza vectorización que convierte el texto en vectores numéricos que pueden ser utilizados por estos algoritmos para entrenar modelos. En este paso se agrega el parametro ngram para analizar las palabras que se escriben juntas y poder obtener un mejor significado de estas.
8. **Balanceo de clases:** Al ver que la columna clases tenía una cantidad de datos desbalanceada. Realizar un balanceo es de gran importancia para evitar sesgos y mejorar la capacidad predictiva. De esta manera se asegura que el modelo aprenderá patrones equitativamente.

Se evaluaron los datos preparados para cada algoritmo utilizando tanto lematización como stemming, seguido de la vectorización para el modelado de los algoritmos. No obstante, se observó que para los tres algoritmos, el stemming con vectorización es más efectivo que la lematización, especialmente en lenguas flexivas como el español. Además, las tres técnicas se implementaron en el cuaderno (notebook) y al ejecutarlo con cada una de ellas, se identificó efectivamente que se obtienen mejores métricas de evaluación al utilizar el stemming y la vectorización.

La vectorización se llevó a cabo utilizando el modelo TF-IDF (Frecuencia de Término - Frecuencia Inversa de Documento), que se considera más adecuado para esta aplicación empresarial. Este modelo asigna un peso a cada palabra según su frecuencia o rareza, lo que la clasifica como más o menos informativa.

3 Modelado y evaluación

Se realizó la aplicación de 3 algoritmos distintos para el aprendizaje automático descrito anteriormente.

Método	**Descripción y justificación de su uso**
Support Vector Machine (SVM)	Las SVM son algoritmos de aprendizaje supervisado utilizados principalmente para la clasificación. Son eficaces en la clasificación de textos, como las reseñas de hoteles, debido a su capacidad para manejar eficazmente conjuntos de datos de alta dimensionalidad. Al ser un modelo de margen máximo, las SVM tienden a generalizar bien y pueden manejar de manera efectiva la separación no lineal entre las clases, lo que es beneficioso para capturar la complejidad de los sentimientos expresados en las reseñas.

Random Forest	Los bosques aleatorios son un conjunto de árboles de decisión que se entrenan con diferentes subconjuntos del conjunto de datos y luego combinan sus resultados para realizar predicciones. Son robustos frente al sobreajuste y funcionan bien con conjuntos de datos grandes y complejos, como las reseñas de hoteles que pueden contener una gran cantidad de características y ruido. Los bosques aleatorios también pueden manejar datos categóricos y numéricos sin necesidad de escalado, lo que los hace adecuados para trabajar con reseñas que pueden contener diversos tipos de información.
Logistic Regression	A pesar de su nombre, la regresión logística se utiliza comúnmente para problemas de clasificación binaria, como determinar si una reseña es positiva o negativa. Es un modelo simple pero efectivo que estima la probabilidad de que una instancia pertenezca a una clase particular. Para el análisis de sentimientos, la regresión logística puede ser útil para proporcionar una interpretación intuitiva de las características más influyentes en la predicción de sentimientos positivos o negativos. Además, es computacionalmente eficiente y puede servir como un buen punto de referencia para comparar con modelos más complejos.

3.1 Evaluación Cuantitativa

En cuanto a la evaluación de los modelos, se emplearon diversos métodos:

1. **Precisión (Accuracy):** Esta métrica indica la proporción o porcentaje de registros que fueron clasificados correctamente por el modelo.

2. **Precisión por Clase (Macro Precisión):** Este enfoque calcula un promedio de la precisión del modelo para cada categoría de clasificación. Es útil para determinar si el modelo es eficiente en la clasificación general de textos entre múltiples categorías, o si su eficacia se limita a clasificar específicamente en una sola calificación.

3. **Recall o Sensibilidad:** Se refiere a la tasa de "verdaderos positivos", es decir, qué proporción de los textos fueron correctamente asignados a su categoría correspondiente de calificación obtenido. Esta métrica es complementaría a la precisión, ya que un modelo puede ser muy sensible pero poco preciso.

4. **Puntaje F1:** Teniendo en cuenta que la precisión y la sensibilidad son métricas complementarias, el puntaje F1 las incorpora a ambas, ofreciendo un promedio ponderado de ambas métricas para proporcionar una medida general del rendimiento del modelo.

También se construye para los modelos la matriz de confusión, que permite visualizar los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para ver si el modelo está cometiendo algún error específico (e.g. si es propenso a caer en falsos positivos). Se muestran a continuación los resultados de esas métricas para cada uno de los modelos. En el notebook se pueden evidenciar gráficas con estas métricas.

Método	**Evaluación cuantitativa**
Support Vector Machine (SVM)	Precisión: 49.87% Accuracy: 50.77% Recall: 50.77% F1 score: 49.66%
Random Forest	Precisión: 35.76% Accuracy: 37.99% Recall: 37.99% F1 score: 34.46%
Logistic Regression	Precisión: 48.85% Accuracy: 49.31% Recall: 49.31% F1 score: 49.03%

Se evidencia entonces que el Support Vector Machine (SVM) es el modelo con mejores indicadores de evaluación cuantitativa es por eso que en el siguiente apartado de este documento se recomienda hacer uso de ese modelo al negocio.

4 Resultados

En esta primera etapa se esperaba encontar un modelo que permitiera clasificar apartados de texto en una de las 5 posibles categorías (o clasificaciones de las reseñas). De los modelos implementados, se recomienda usar el modelo de Support Vector Machine (SVM) esto se debe a que presenta los mejores indicadores al momento de la evaluación. Los sitios con calificaciones bajas podrían beneficiarse de estrategias de mejora basadas en las características identificadas en las reseñas, mientras que aquellos con altas calificaciones podrían fortalecer aún más sus puntos fuertes para atraer a más turistas.

Se destaca el enfoque en las características importantes identificadas por el modelo Random Forest, a pesar de no haber sido el modelo seleccionado como el mejor. Estas características proporcionan información valiosa sobre las palabras clave que influyen significativamente en el análisis de sentimientos. Las tres raíces de palabras más importantes, según el modelo Random Forest, fueron "mal", "excelent" y "mas". Estas palabras sugieren que en las reseñas analizadas, se mencionan aspectos como la calidad del servicio o producto ("excelent"), la comparación con otras opciones ("mas"), y la experiencia general ("al"). Por otro lado, se observan otras raízes de importancia como "suci", "com" y "servici", estas nos indican que la suciedad, la comida y el servicio son de gran importancia.

4.1 Resultados estadística

Al reunirnos con el grupo de estadística se pudo llegar a una conclusión con respecto al resultado mejor obtenido que fue de 0.4746.

Referencia: p = 0.99 Experimental: $\hat{p} = 0.5077$

n: 4662 l = 0.5

1. Requisitos:
   Muestra aleatoria? Asumimos que sí
   $n_p \geq 5 => 4662 \cdot 0.99 = 4615,38 \geq 5$
   $n_q \geq 5 => 4662 \cdot 0.01 = 46,42 \geq 5$

2. Hipótesis:
   $h_0 : p = 0.99$
   $h_a : p < 0.99$ (dado que $\hat{p}$ es 0.5077)
   Entonces se hace una "Prueba de cola izquierda"

3. Estadístico de prueba:
   $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} = \frac{0.5077-0.99}{\frac{0.99 \cdot 0.01}{4662}}$ $z = -0.0010397$

4. Calculadora de distribución normal estandár:
   Valor $p = 0.49959$

5. Conclusion:
   Rechazo H0 Rechazo Ha
   Valor $p < l <$ Valor p
   Valor $p(0,49959) <$ significacia $l(0.5)$

Se logró llegar a la conclusión que tenemos evidencia significativamente estadística para afirmar que la probabilidad experimental es más acertada que la de referencia.

5 Mapa de actores relacionado con el producto de datos creado

Actor	Beneficio	Riesgo
Ministerio de Comercio, Industria y Turismo	Obtención de información detallada sobre las características de los sitios turísticos en Colombia para diseñar políticas y estrategias turísticas.	Dependencia excesiva de los datos y resultados del modelo analítico, limitando la consideración de otros factores importantes en la promoción turística.
Asociación Hotelera y Turística de Colombia	Acceso a insights sobre las preferencias y opiniones de los turistas para mejorar servicios y atraer más clientes.	Interpretación errónea de los resultados del análisis, llevando a decisiones estratégicas poco acertadas en la promoción y gestión de destinos turísticos.

Cadenas Hoteleras	Identificación de oportunidades para adaptar ofertas y servicios según preferencias de los turistas, aumentando la satisfacción del cliente y ocupación hotelera.	Sobrevaloración o subestimación de aspectos turísticos debido a la limitación del análisis a datos específicos, conduciendo a decisiones comerciales erróneas.
Hoteles Pequeños	Acceso a información valiosa sobre características y preferencias de turistas en diferentes destinos para mejorar su oferta y competir de manera más efectiva.	Dependencia exclusiva del modelo analítico sin considerar otros factores contextuales locales, conduciendo a decisiones inadecuadas en la gestión y promoción turística.
Turistas	Recepción de recomendaciones más personalizadas sobre destinos turísticos para tomar decisiones informadas y disfrutar de experiencias más satisfactorias.	Pérdida de autenticidad y singularidad de destinos turísticos debido a la estandarización de ofertas y servicios impulsada por estrategias basadas en análisis de datos.

6 Trabajo en equipo

Roles de cada integrante en el equipo y aporte:

Integrantes	**Participación y roles adoptados**
María Alejandra Estrada García	Roles: Líder de proyecto, Líder de analítica. Acumulado de número de horas dedicadas al proyecto: aproximadamente 12 horas de trabajo neto. Algoritmo(s) trabajados: Support Vector Machine Retos enfrentados en el proyecto y formas planteadas para resolverlos: Gestionar eficazmente el tiempo y la Ingeniería de Sistemas y Computación carga de trabajo, implementé una estructura de gestión del tiempo más efectiva y me rodeé de un entorno de trabajo inspirador. Puntos repartidos: 33.33/100.

Santiago Martínez Novoa	Roles: Líder de negocio. Acumulado de número de horas dedicadas al proyecto: aproximadamente 12 horas de trabajo neto. Algoritmo(s) trabajados: Logistic Regression Retos enfrentados en el proyecto y formas planteadas para resolverlos: La gestión del tiempo. Para abordar este reto, implementé una planificación detallada, asignando tareas específicas a intervalos regulares y estableciendo plazos realistas. Puntos repartidos: 33.33/100.
Marilyn Stephany Joven Fonseca	Roles: Líder de datos. Acumulado de número de horas dedicadas al proyecto: aproximadamente 12 horas de trabajo neto. Algoritmo(s) trabajados: Random Forest. Los retos estuvieron relacionados con la gestión del tiempo fue crucial en este proyecto. Para manejar esta situación, opté por una estrategia de planificación meticulosa. Esta implicó la asignación de tareas específicas en intervalos regulares, así como la fijación de plazos alcanzables y realistas. Puntos repartidos: 33.33/100.

Reuniones de grupo:

- **Reunión de lanzamiento y planeación:** Para definir roles y forma de trabajo del grupo. Se genera lluvia de ideas sobre la forma de resolver el proyecto.

  Día: 20 de marzo, 2024 Hora: 3:00pm

- **Reunión de ideación:** Una vez se han explorado los datos del proyecto, la reunión de ideación busca definir la organización/empresa/institución y el rol dentro de ella, que se beneficia de la solución analítica que van a desarrollar

  Día: 22 de marzo, 2024 Hora: 5:00pm

- **Reuniones de seguimiento:** Se recomienda mínimo una reunión de seguimiento semanal corta. También pueden ser correos de avance según lo defina el grupo. Pueden tener un tablero de control de las tareas, utilizando herramientas como Trello.

  1. Día: 30 de marzo, 2024 Hora: 3:00pm 2. Día: 3 de abril, 2024 Hora: 7:00 am 3. Día: 5 de abril, 2024 Hora: 5:00 pm 4. Día: 6 de abril, 2024 Hora: 10:00 am

- **Reunión de finalización:** Para consolidar el trabajo final, verificar el trabajo del grupo y analizar los puntos a mejorar para la siguiente etapa del proyecto.

  Día: 6 de abril, 2024 Hora: 4:00pm

References

[1] Statista. (2023, 15 octubre). *Colombia: número de hoteles y establecimientos similares 2005-2021.* https://es.statista.com/estadisticas/1018140/evolucion-anual-del-numero-de-hoteles-y-establecimientos-similares-en-colombia/

[2] Elastic. (s. f.). *¿Qué es el procesamiento de lenguaje natural (NLP)? — Una guía completa del NLP.* https://www.elastic.co/es/what-is/natural-language-processing

[3] MATLAB & Simulink. (s. f.). *Support Vector Machine (SVM).* https://la.mathworks.com/discovery/support-vector-machine.html

[4] Fontalvo, W., Mendoza Vega, L., Diaz Solano, B., Hereira, M. J., Torres, D., Diago, F., Martinez, J. (2022). *Caracterización y uso de las redes sociales en las empresas del sector hotelero en la ciudad de Barranquilla - Colombia.* Saber, Ciencia Y Libertad, 17(2), 238–256. https://doi.org/10.18041/2382-3240/saber.2022v17n2.9278

[5] Camilo, Y., & Betancur, F. (2020) *FACTORES DETERMINANTES PARA LA RESERVA DE HOTELES EN LÍNEA EN COLOMBIA. UNA MIRADA DESDE EL MODELO DE ACEPTACIÓN TECNOLÓGICA (TAM) AJUSTADO.* TURPADE. Turismo, Patrimonio Y Desarrollo, 12. https://revistaturpade.lasallebajio.edu.mx/index.php/turpade/article/view/37

[6] DAC. (2019, April 16). *a importancia de la reseñas para los hoteles.* DAC España. DAC España. https://www.dacgroup.com/es/blog/la-importancia-de-la-resenas-para-los-hoteles/

[7] Statista. (2021). *Número de hoteles en Colombia 2021.* https://es.statista.com/estadisticas/1018140/evolucion-anual-del-numero-de-hoteles-y-establecimientos-similares-en-colombia/