Integrantes:

Santiago Sinisterra – 202022177

Lina Ojeda - 202112324

Ana Sofía Padilla Daza - 202021748

Proyecto 1, Etapa 1 – Analítica de textos

Inteligencia de Negocios

Contenido

Entendimiento del negocio y enfoque analítico ..... **¡Error! Marcador no definido.**

Descripción del enfoque analítico para alcanzar los objetivos del negocio ..... **¡Error! Marcador no definido.**

Entendimiento y preparación de los datos ..... **¡Error! Marcador no definido.**

Modelado y evaluación ..... **¡Error! Marcador no definido.**

Resultados ..... **¡Error! Marcador no definido.**

Mapa de actores relacionado con el producto de datos creado. ..... **¡Error! Marcador no definido.**

Trabajo en equipo ..... **¡Error! Marcador no definido.**

Entendimiento del negocio y enfoque analítico

Enunciado:

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.

Descripción del enfoque analítico para alcanzar los objetivos del negocio

El **objetivo** del equipo de trabajo es construir un modelo analítico que permita realizar la calificación automática de nuevas reseñas, con un alto nivel de precisión y de sensibilidad (recall) en base al análisis de reseñas anteriores. Como **criterios de éxito** se establecieron dos de gran importancia: El primero es que el modelo escogido al final sea capaz de recibir una reseña en formato de texto y clasificarla en que calificación caería (entre 1 y 5, siendo 1 muy malo y 5 muy bueno). Como segundo criterio se busca que el modelo exhiba una probabilidad mayor al 80% de que la clasificación asignada sea correcta.

Este proyecto podría beneficiar a la industria del turismo en general, pero más específicamente a los dueños de hoteles en Colombia. Según la información aportada por Encuesta Mensual de Alojamiento (EMA) del DANE fue posible identificar como la ocupación hotelera disminuyó del 2022 al 2023 en un 3,2% (Economía, 2024). Esto da a entender que actualmente la industria hotelera colombiana se está enfrentando a varios desafíos. Por eso este proyecto podría tener un gran impacto en identificar los aspectos que deberían mejorar los hoteles para aumentar su ocupación, asimismo, este proyecto podría aportarle a los dueños de los hoteles que buscan construir nuevos alojamientos, información sobre qué deberían tener estos para poder atraer más clientes.

Oportunidad / Problema negocio	El objetivo del negocio es poder determinar una clasificación a un comentario con respecto al contenido del anterior. Con esto en mente, se podría considerar que un criterio de éxito desde el punto de vista del negocio sería que al 85% de los datos de prueba se les asigne una clasificación correspondiente por medio del uso del modelo de machine learning entrenado.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	El requerimiento desde el punto de vista de aprendizaje automático es entrenar un modelo de Machine Learning de forma que pueda recibir un comentario y respectivamente determinar la calificación esperada con respecto al contenido del anterior. Esto se realizará utilizando un preprocesamiento de datos ya calificados con el modelo de bolsa de palabras, utilizando en específico un modelo de frecuencias absolutas. Además, se utiliza lematización en el preprocesamiento de los datos para obtener mejores resultados al aplicar los modelos. A lo que seguido, se considera la aplicación de tres algoritmos: clasificación (árbol de decisión o random forest), regresión logística y Naive Bayes. Estos siendo aplicados a varios conjuntos de datos preprocesados, para poder, entre los tres, determinar cuál es el que tiene menor error y con qué set de datos da mejores métricas e implementarlo en los datos sin clasificación suministrados.

Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización en general sería la industria hotelera colombiana que se podría beneficiar de automatizar la clasificación de los textos de varias maneras. Pero, sobre todo, serviría para acelerar y mejorar el proceso de mejoramiento y retroalimentación. Esto a su vez le sería de gran utilidad a los encargados de garantizar la calidad de los hoteles pues les permitiría entender y analizar de manera más eficiente tanto las cosas que le agradan a los clientes, para poder continuar haciéndolas, como aquellos aspectos de la atención que se deben mejorar.
Contacto con experto externo al proyecto y detalles de la planeación	Contactamos con las estudiantes de estadística: Sarita Garzón Diana Rubio Reunión para compartir resultado de modelo y planear siguiente fase: viernes 29 de marzo. También se plantea una reunión el 8 de abril para revisar los entregables y el enfoque que se va a tomar para la etapa 2.

Entendimiento y preparación de los datos

Realizando una primera aproximación a los datos se pudo identificar que el archivo dado contiene 7875 filas y dos columnas. Estas columnas corresponden a la reseña dada a un hotel y a la clasificación proporcionada en esta reseña. La clase de la reseña puede ser un número entre 1 y 5, siendo 1 una mala calificación y 5 una buena. Se observó que todas las 7875 filas contenían información, esto es, que no contenían espacios vacíos en ninguna de las dos columnas, pero que 71 estaban duplicadas. También, cada reseña tiene una clasificación entre 1 y 5 sin que haya alguna que tenga una mayor o menor al rango esperado.

Con base en el perfilamiento, se analizó la calidad de los datos considerando las dimensiones de unicidad, consistencia, validez, y completitud. Respecto a la unicidad, se consideraron los valores duplicados, se encontró que estos son 71 valores y se eliminaron correspondientemente. De forma respectiva, la consistencia de los datos es asegurada al modificar los datos para que siempre tengan el mismo formato de tokenización esta permite que el texto se divida en unidades coherentes, y se eliminen todos los aspectos considerados necesarios para que las palabras tengan un mismo formato, para que estas sean palabras legibles, sin números, sin puntuación, además se realiza la lematización sobre ellos lo cual reduce las palabras a sus formas base, lo que facilita la comparación y el análisis de los datos.

La validez se garantiza al procesar todos los valores para que sean palabras válidas, excluyendo aquellas que pertenezcan al conjunto de stopwords en español. De esta manera, el procesamiento se asegura de que las palabras utilizadas sean relevantes y no estén dentro de un conjunto considerado no informativo, conocido como stopwords.

Después de este proceso, contábamos ahora con un conjunto de datos relativamente limpio. Al procesar el lenguaje natural, se decidió realizar una vectorización a "bag of words" de los datos. Esto consiste en una representación de la reseña en la que se ignora el orden de las palabras y se considera únicamente la frecuencia con la que cada palabra aparece. Además, luego del proceso anteriormente descrito se optó por eliminar aquellas palabras que aparecen tres veces o menos en el conjunto de datos. Esta decisión se tomó tras considerar que estas palabras, dentro de un conjunto de aproximadamente 18,000 palabras, probablemente no tengan un impacto significativo para proporcionar una contribución relevante al análisis o procesamiento de los datos.

Además, se asegura la completitud de los datos al eliminar todos los comentarios que después de la eliminación de palabras quedaran sin palabras consideradas. Por último, se eliminó palabras que contuvieran números dentro de ellas, ya que consideramos que no eran relevantes para la predicción.

Según lo mencionado. Estas decisiones fueron realizadas:

1. Eliminación de valores duplicados.
2. Eliminación de valores nulos.
3. Eliminación de números.
4. Eliminación de palabras stopwords.
5. Lematización
6. Se aplico la vectorización a bag of words.
7. Eliminación de palabras que solo tengan tres instancias.
8. Eliminación de palabras que incluyen números.

Así mismo, vale la pena resaltar que se revisó en los algoritmos de clasificación los resultados independientes de cada aspecto realizado del preprocesamiento, para así evaluar la validez de este y si da mejores resultados que la sola implementación de los datos tokenizados, lo cual resulto en que los filtros implementados si aportan de forma significativa a las métricas deseadas.

Modelado y evaluación

Después de tener los datos procesados, se realizaron 4 algoritmos de aprendizaje automático con distintos sets de datos. Se hicieron para poder recibir como entrada una reseña en forma de texto para luego clasificarla en alguna clase entre 1 y 5. A pesar de que solo era requerido realizar 3 algoritmos, se tomó la decisión de añadir un 4, siendo este Random Forest. Esto debido que inicialmente el modelo implementado con árboles de decisión no dio los resultados esperados, por lo cual se quiso buscar un algoritmo similar de clasificación que fuera más eficiente. Así se llegó al algoritmo de Random Forest. Para demostrar esté proceso y las diferencias entre los resultados de ambos modelos, se decidió mantener ambos algoritmos en la ejecución. A continuación, se describen los cuatro métodos usados.

Método	**Descripción y justificación de su uso**	
Clasificación	Árboles de decisión	Los árboles de decisión son un método de aprendizaje supervisado. Este crea arboles en los cuales sus nodos hoja representan los resultados posibles y los nodos intermedios representan las decisiones que se toman. Así, dependiendo de distintas decisiones el árbol llega a clasificar los datos en los nodos hoja. Los árboles de decisión tienen la ventaja que no necesitan una preparación de datos tan detallada. Además, a la hora de representar el proceso de clasificación o regresión, los árboles proveen una forma fácil de visualizarlo que puede resultar útil para muchos clientes. Y aunque no es el algoritmo más detallado se eligió compararlo con su variante Random Forest. (IBM, s.f.)
	Random Forest	Este algoritmo está basado en árboles de decisión. A pesar de la efectividad de los árboles de decisión estos a veces pueden tender a estar sesgados o sobreajustar. Por esto, al tomar varios árboles, cada uno con datos distintos del conjunto de entrenamiento, y combinar sus resultados se pueden obtener resultados más precisos y robustos, que es exactamente lo que hace el algoritmo Random Forest. Este algoritmo tiene tres principales ventajas por las cuales resulta útil a la hora de clasificar textos: Primero, en relación con lo mencionado anteriormente, el algoritmo reduce el riego de sobreajuste. Segundo, puede aportar alta precisión incluso si hay datos faltantes o si hay una gran cantidad de datos. Y tercero, es más simple de entrenar comparado con otros algoritmos que dan resultados similares. Todas estas características hacen que resulte muy útil para la tarea planteada. (Espinosa-Zúñiga, 2020)
Logistic regression	Logistic regression es un tipo de modelo estadístico común mente utilizado para problemas de clasificación, este siendo un método de aprendizaje automático se utiliza para predecir la ocurrencia de un evento en términos binarios es decir opciones como si o no, falso o verdadero, bueno o malo, entre otras. Este método transforma linealmente las variables predictoras, como las palabras utilizadas en una reseña, asignándoles pesos que representan su importancia para la predicción esto mediante una función logística no lineal. Con esto obtenemos valores de predicción de 1 y 0, con estos valores se revisa un umbral que las clasifica. Este método es usado en muchas áreas distintas como en medicina, deportes o incluso en la bolsa, esto pues es fácil de comprender incluso para aquellos que no tienen conocimiento previo en Machine Learning y sus resultados son más fáciles de presentar. Esto resulta sumamente importante para este proyecto.	

Naive Bayes	Naive Bayes es un modelo se basa en el teorema de probabilidad condicional de Bayes. Presume de antemano independencia condicional entre los textos de ahí el nombre *naive.* Este clasificador es fácil de implementar, es eficiente computacionalmente y funciona muy bien tanto en sets grandes como pequeños. Además, se caracteriza por arrojar resultados muy buenos, sobretodo en modelos de procesamiento de texto en lenguaje natural, lo cual es sumamente apropiado para este contexto de negocio.

Evaluación cuantitativa de los modelos:

Para la evaluación cuantitativa se usaron los criterios de Accuracy, recall y F1 principalmente para determinar la efectividad de cada uno. La métrica de Accuracy o presición indica qué proporción o porcentaje de los registros fueron clasificados de forma correcta. El Recall o sensibilidad mide la tasa de "verdaderos positivos" clasificados correctamente, es decir, qué proporción de los textos fueron asignados su clase correcta. Se complementa con la precisión, pues puede pasar que un modelo sea muy sensible pero poco preciso. Y finalmente se usó el F1 score que contempla tanto precisión como sensibilidad.

Además, se construyó una matriz de confusión para cada modelo que permite visualizar lo expresado en estas estadísticas a través de la cantidad total de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Esto también ayuda a comprender si alguno de los modelos es más propenso a producir algunos valores erróneos en específico.

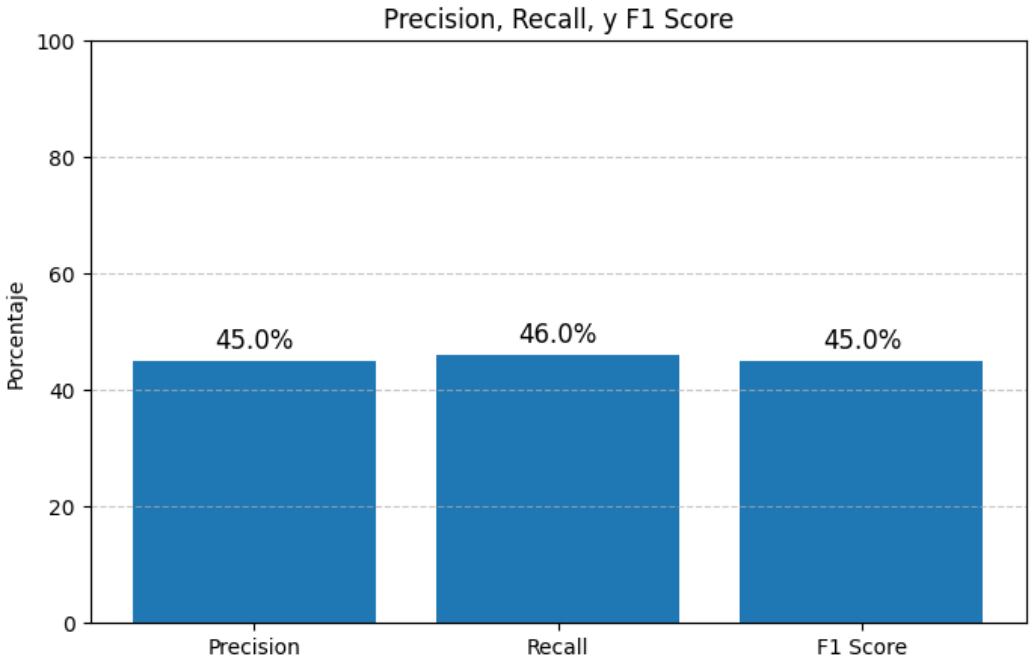
Métricas de los modelos:

Método	**Evaluación cuantitativa**
Árboles de decisión	Accuracy: 36% Recall: 37% F1: 34%
Random Forest	Accuracy: 45% Recall: 46% F1: 45%
Logistic Regression	Accuracy: 100% Recall: 100% F1: 100%
Naive Bayes	Accuracy: 55.15% Recall: 44.64% F1: 26%

En base a estas métricas se concluyó que el mejor modelo para la tarea es Random Forest con todos los filtros descritos anteriormente implementados y sin optimización de parámetros. Esto pues, a pesar de que logistic regression fue el modelo con mejores métricas creemos que esto se debe a que está overfitted, por lo cual elegimos el segundo mejor algoritmo, siendo este Random Forest.

Resultados

El resultado obtenido en esta primera etapa del proyecto es un modelo predictivo que busca clasificar distintas reseñas en formato de texto en una categoría entre 1 y 5 para determinar que tan positivas o negativas son. Como se dijo anteriormente, este modelo terminó siendo el modelo basado en el algoritmo Random Forest que presenta los siguientes indicadores:



Estos demuestran que a pesar de que el modelo no es perfecto, si resultaría útil para la organización para determinar la clasificación de aquellas reseñas que no tengan. Como estos resultados no fueron tan buenos como se esperaba se hizo un análisis de posibles problemáticas. Una de ellas fue la forma de preparar los datos, que pudo incluir palabras no importantes o, al contrario, eliminar palabras muy relevantes para las reseñas. Sin embargo, al darnos cuenta de esto se trato de mejorar esta preparación lo más posible. La otra teoría fue que muchas reseñas en los datos de entrenamiento no estaban escritas acorde a sus calificaciones. Por ejemplo, esta reseña: “Lo mejor era la limonada. Me gusto la comida de todo el mundo y era sosa y un poco frío” que tiene clasificación de 2, dando a entender que sería mala, sin embargo, palabras como mejor y gusto pueden hacer que el modelo intuyera que esta reseña tendría una clasificación más alta.

También es importante añadir que en el análisis de las palabras no se tuvieron en cuenta caracteres como emoticones o puntuaciones, que son muchas veces usadas para expresar emociones, por lo que esto también pudo afectar el análisis. Además, creemos que si se pudieran agrupar algunas palabras esto mejoraría el modelo, palabras como lugar, comida o servicio se podrían juntar con el adjetivo que las defina.

En base al modelo también se identificaron varias palabras que fueron relevantes a la hora de caracterizar las reseñas, entre estas las más importantes fueron: excelente, buen, hotel, habitación, malo, lugar, decir, mal, comida, poder, si, pésimo, servicio, bien, sucio, hacer, él, ver, restaurante, ir. Estas palabras dan cuenta de los resultados dados, pues mientras que muchas palabras tienen la importancia adecuada, como buen, mal, excelente, bien, pésimo, otras no son muy relevantes para clasificar las reseñas como: él o si.

A partir de estos resultados se podrían realizar recomendaciones a la empresa para utilizar los resultados. Por un lado, se recomendaría usar el modelo para poder encontrar reseñas con bajas clasificaciones para poder identificar los principales problemas que presentan los hoteles y que se puedan asignar recursos a corregirlos. Por otro lado, se recomendaría que se hiciera lo mismo mencionado anteriormente, pero identificando las reseñas con clasificación de 5 para poder mantener los aspectos positivos. Finalmente, si una cadena de hoteles deseara usar el modelo, podría aplicarlo a varios hoteles en diferentes lugares y aplicar lo identificado en las reseñas positivas de algunos para mejorar la atención en las sede que no tengan un promedio de calificación tan elevado.

Mapa de actores relacionado con el producto de datos creado.

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Directivas de un hotel	Usuario-Cliente	Ayuda a mejorar la calidad de los servicios del hotel. Con esto se pueden mejorar las ganancias del hotel y proporcionarle información a las directivas de la satisfacción de sus clientes.	Si el modelo no tiene un desempeño rápido o correcto, esto puede hacer que las directivas asuman que sus inversiones en el modelo fueron perdidas o que al confiar demasiado en los resultados del modelo se obtenga información incorrecta sobre la calidad del hotel, perjudicando así las ventas y los alojamientos.

Área de gestión de calidad de hotel	Usuario-Cliente	Facilita la recolección de datos sobre la satisfacción de los usuarios frente al hotel. Con esta información el área de gestión de calidad del hotel podría saber en qué aspectos de el hotel deberían enfocarse en mejorar o eliminar.	Si el modelo no tiene un buen desempeño el área de gestión de calidad podría asumir que hay problemas en donde no los hay o podría confundirse a la hora de analizar las reseñas clasificadas en lo que quiere el cliente.
Área de financias del hotel	Financiador	Con la clasificación, se puede hacer una asignación eficiente de recursos financieros hacia proyectos específicos relacionados con los problemas del hotel	Si el modelo no tiene un buen desempeño existe el riesgo que el área de finanzas decida invertir en ares que no son prioritarias para la organización o incluso cortar gastos en áreas que, si lo son, lo cual le resultaría costoso a la empresa.
Huésped del hotel	Beneficiado	Gracias a la clasificación implementada el huésped podrá ser atendido de mejor manera, además sus quejas o felicitaciones acerca de los servicios del hotel serán tomadas en cuenta con mayor facilidad	Si el hotel afirmara haber tomado en cuenta los comentarios de los huéspedes anteriores, pero estos estuvieran mal clasificados, los nuevos huéspedes podrían asumir que el hotel no toma en cuenta sus opiniones y no volver a hospedarse ahí.

Trabajo en equipo

Integrate 1:

Integrate	Rol	Algoritmos trabajados	Retos enfrentados	Solución a los retos
Santiago Sinisterra	Líder de datos	Arboles de decisión y Random Forest	Manejo del tiempo de las ejecuciones, saber que filtros eran los adecuado para usar.	Para el manejo del tiempo de las educaciones no se pudieron realizar muchas pruebas y nos conformamos con las realizadas y para la decisión sobre los filtros adecuados se realizó una evaluación de los datos con cada filtro y su eficiencia.

Integrante 2:

Integrate	Rol	Algoritmos trabajados	Retos enfrentados	Solución a los retos
Lina Ojeda	Lider de negocio	Logistic Regression	Entendimiento de cómo funcionaba el algoritmo, saber que variables aceptaba y cuales no, y manejo de tiempo.	Leer documentación acerca de logistic regression, revisar junto con el profesor el error y encontrar una solución pertinente. Aunque no se pudieron realizar las pruebas esperadas por el tiempo de ejecución que tomaba se decidió trabajar con lo que teníamos.

Integrante 3:

Integrate	Rol	Algoritmo trabajado	Retos enfrentados	Solución a los retos
Ana Sofia Padilla Daza	Lider de proyecto y Líder de analítica	Naive Bayes	El manejo del tiempo, el entendimiento completo del problema.	Para poder entender mejor el problema y los pasos a seguir para desarrollar el proyecto pedí ayuda tanto a monitores y profesores como a mis compañeros de equipo. Para optimizar el manejo del tiempo y la carga de trabajo realicé una planeación de lo que iba a hacer cada día y utilicé la técnica Pomodoro para poder lograr los objetivos diarios.

Repartición de puntos entre los integrantes:

Integrante	Puntos sobre 100	Horas dedicadas al proyecto en total
Santiago Sinisterra	34	24 horas
Lina Ojeda	33	20 Horas
Ana Sofía Padilla Daza	33	20 horas

Reuniones:

Numero de la reunión	Propósito	Fecha	Duración	Asistentes
1	Entender el problema.	25/03/2024	2 horas	Lina Ojeda Santiago Sinisterra Ana Sofia
2	Realizar preprocesamiento de los datos	26/03/2024	8 horas	Lina Ojeda Santiago Sinisterra Ana Sofia
3	Realizar modelación de los datos	31/03/2024	4 horas	Lina Ojeda Santiago Sinisterra Ana Sofia
4	Finalización del documento y conclusión de detalles	06/04/2024	2 horas	Lina Ojeda Santiago Sinisterra Ana Sofia

Referencias

1. Espinosa-Zúñiga, Javier Jesús. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería, investigación y tecnología, 21(3), 00002. Epub 02 de diciembre de 2020.https://doi.org/10.22201/fi.25940732e.2020.21.3.022
2. Economía, R. (2024, 19 enero). Hoteles tuvieron un 2023 de caídas en ocupación e ingresos: ¿qué dicen las cifras? ELESPECTADOR.COM. https://www.elespectador.com/economia/hoteles-tuvieron-un-2023-de-caidas-en-ocupacion-e-ingresos-que-dicen-las-cifras-noticias-colombia/
3. ¿Qué es un árbol de decisión? | IBM. (s. f.). https://www.ibm.com/es-es/topics/decision-trees
4. ¿Qué es la regresión logística? | IBM. (s. f.). https://www.ibm.com/mx-es/topics/logistic-regression
5. Raschka, Sebastian, and Vahid Mirjalili. Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2. 3rd ed. Birmingham: Packt Publishing, Limited, 2019. Print.