Universidad de los Andes: Proyecto 1 – Inteligencia de Negocios

Jeronimo Vargas Rendon - 202113305

Juan Manuel Pérez - 202021827

Felipe Núñez - 202021673

Turismo de los Alpes

Tabla de Contenidos

1. Contexto 2
1.1. Objetivos 2
1.2. Criterios de Éxito 2
1.3. Impacto en Colombia: 2
1.4. Enfoque Analítico 2
2. Entendimiento y preparación de los datos 3
2.1. Perfilamiento y análisis de calidad 3
2.2. Tratamiento de los datos 4
2.2.1. Limpieza de datos 4
2.2.2. Tokenización 4
2.2.3. Normalización 5
2.2.4. Transformación 5
3. Modelado y evaluación 5
3.1. Modelos Term Frequency Inverse Document Frequency 5
3.1.1. Regresión logística 5
3.1.2. Multinomial Naive Bayes 6
3.1.3. Random Forest 6
3.2. Count Vectorizer 6
3.2.1. Regresión Logistica 6
3.2.2. Multinomial Naive Bayes 6
3.2.3. Random Forest 7
3.3. Count Vectorizer Dummy 7
3.3.1. Regresión Logistica 7
3.3.2. Multinomial Naive Bayes 7
3.3.3. Random Forest 8
3.4. Selección del modelo 8
4. Resultados 8
5. Mapa de actores 9

6. Trabajo en equipo 9

1. Contexto

1.1. Objetivos

- **Mejorar la Experiencia Turística:** Utilizar las reseñas y calificaciones para identificar los elementos más valorados por los turistas en los sitios altamente recomendados.
- **Incrementar la Competitividad:** Analizar características de sitios con bajas calificaciones para implementar mejoras, aumentando así su atractivo y competitividad.
- **Promoción y Recomendación:** Desarrollar un sistema de recomendación basado en datos para guiar a turistas locales e internacionales hacia los sitios que más se alinean con sus preferencias.

1.2. Criterios de Éxito

- **Precisión en la Calificación:** El modelo analítico debe calificar las reseñas con alta precisión, reflejando fielmente la percepción y experiencia de los turistas.
- **Impacto en la Decisión del Turista:** Las recomendaciones y calificaciones proporcionadas deben influir positivamente en las decisiones de los turistas, incrementando la visita a sitios recomendados.
- **Mejora Continua:** El sistema debe permitir la identificación constante de oportunidades de mejora en los servicios y experiencias ofrecidas por los sitios turísticos.

1.3. Impacto en Colombia:

- **Crecimiento Económico:** Al mejorar la experiencia turística y aumentar la competitividad de los sitios turísticos, se espera un aumento en el turismo que contribuya al crecimiento económico local.
- **Reputación Internacional:** Mejorar la calidad y la percepción de los sitios turísticos colombianos puede fortalecer la posición de Colombia como destino turístico a nivel mundial.
- **Desarrollo Sostenible:** Fomentar un turismo basado en la calidad y la sostenibilidad puede contribuir al desarrollo sostenible de las comunidades locales, preservando al mismo tiempo los recursos naturales y culturales del país.

1.4. Enfoque Analítico

El conjunto de datos que consiste en reseñas sobre hoteles y los lugares turísticos colombianos. Cada reseña está asociada a una calificación que refleja la experiencia del turista. La variable de interés en este análisis es la calificación (“Class”) asignada a cada reseña, la cual se

considera nuestra variable objetivo (variable Y). Este conjunto de datos etiquetado nos permite utilizar técnicas de aprendizaje supervisado para entrenar un modelo que pueda predecir la calificación de los sitios turísticos. Para realizar el análisis de las reseñas sobre hoteles y lugares turísticos colombianos, primero necesitamos realizar algunas transformaciones en los datos para que sean adecuados para su procesamiento. Es necesario procesar el texto que consiste en tokenización, eliminación de stopwords, lematización y vectorización. Luego se divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba para poder aplicar los modelos de aprendizaje automático supervisado de Random Forest, Regresión Logistica y Naive Bayes. Estos son los algoritmos que vamos a utilizar para predecir la calificación de las reseñas en nuestro conjunto de datos y seleccionaremos el mejor basado en sus métricas como lo son la precisión, recall y F1-Score.

2. Entendimiento y preparación de los datos

2.1. Perfilamiento y análisis de calidad

De entrada, tenemos que se encuentran 7875 observaciones, y hay dos variables. La primera es la variable Review que es una cadena de cáracteres y hace referencia a el comentario de los viajeros. La segunda variable es Class, esta variable es nuestra variable objetivo pues esta representa la satisfacción del viajero.

Análisis de Dimensión de calidad de datos:

- Unicidad:

Para verificar la unicidad se calculan la duplicación de las filas.

Duplicados	71

Hay 71 observaciones duplicadas, es necesario eliminar las filas repetidas y dejar solo una observación. Ya que dejarlas incrementaría los efectos de algunas palabras dada su repetición en el modelo.

- Completitud:

Para verificar la completitud se calculan el número de observaciones faltantes por variable.

Variable	Número de observaciones faltantes
Review	0
Class	0

No era necesario hace ningún tratamiento pues todas las variables estaban completas.

- Validez:

Para verificar la validez, se conocía que la variable objetivo Class únicamente podía tomar valores entre [1,5], para esto se calculó el valor mínimo y el máximo de esta variable.

Class

Mínimo	1
Maximo	5

Con esta información podemos concluir que el formato de la variable Class cumplía con las reglas de negocio y no era necesario realizar ninguna transformación.

- Consistencia:

2.2. Tratamiento de los datos

2.2.1. Limpieza de datos

Para este proceso de limpieza de datos es necesario hacer una limpieza de las palabras que no agregan valor. Para esto se definió la función remove_ noise que se encarga de eliminar ciertos tipos de ruido o elementos no deseados del texto antes de continuar con el proceso de limpieza.

1. Elimina todas las etiquetas HTML presentes en el texto. Esto es importante si estás trabajando con texto extraído de páginas web o documentos HTML, ya que las etiquetas HTML no proporcionan información útil para el análisis de texto y pueden interferir con los procesos de limpieza y análisis posteriores.

2. Elimina todos los números del texto. Al observar algunas de los Review muchos de los números que se encontraban, hacen referencia a fechas o horas. Por esta razón, consideramos que los números no agregaban valor a estos comentarios y se eliminaron los números de los comentarios.

3. Elimina todos los caracteres especiales excepto los espacios en blanco y los caracteres alfabéticos (incluyendo letras con tilde y la ñ). Esto es útil para eliminar signos de puntuación, símbolos y otros caracteres que no aportan significado semántico al texto. Sin embargo, conserva los espacios en blanco para preservar la estructura del texto.

4.Reemplaza múltiples espacios en blanco consecutivos por un solo espacio en blanco. Esto ayuda a normalizar la cantidad de espacios en el texto, lo que facilita el procesamiento posterior.

5. Convierte todo el texto a minúsculas. Esto es importante para garantizar la consistencia en el texto y evitar problemas de diferenciación entre mayúsculas y minúsculas durante el análisis.

La otra función que se definió es clean_text, que se encarga de las "stopwords". Las "stopwords" son palabras que se filtran del texto durante el análisis de procesamiento del lenguaje natural (NLP) porque no aportan un significado contextual importante para el análisis. La función específicamente tokeniza el texto en palabras y luego filtra aquellas que no son stopwords ni caracteres alfabéticos, reconstruyendo el texto sin las stopwords.

2.2.2. Tokenización

En la tokenización se observa que se aplica la función de preprocesamiento que se explicó anteriormente. Como resultado de esta tokenización preprocesada, se obtienen únicamente las palabras que cuentan con un sentido semántico. Esto significa que las palabras resultantes después del preprocesamiento son aquellas que tienen relevancia para el análisis de texto.

2.2.3. Normalización

El objetivo principal de la lematización es reducir las palabras a su forma canónica para facilitar el análisis semántico. Para esto se utilizó la librería de Stanza desarrollada por el grupo de investigación de la Universidad de Stanford. En la siguiente función, Stanza realiza la lematización de las palabras en español. Después de procesar el texto, Stanza identifica y reemplaza cada palabra con su lema correspondiente.

2.2.4. Transformación

Para poder realizar un modelo análisis de textos es necesario representar el texto numéricamente para poder aplicar los algoritmos de aprendizaje automático. Hay diferentes transformaciones posibles, se probaron 3:

- Term Frequency Inverse Document Frequency: Es una medida estadística que evalúa la importancia de una palabra en el Review. Creando una matriz por cada Review y cada termino.

- Count vectorizer: Cuenta la frecuencia de ocurrencia de cada palabra en cada Review. Por esta razon construye una matriz de ocurrencias.

- Count Vectorizer Dummy: Cuenta las ocurrencias de palabras, simplemente marca la presencia o ausencia de una palabra en un Review utilizando valores binarios (1 para presencia, 0 para ausencia).

3. Modelado y evaluación

Dado que se realizaron 3 transformaciones se corrieron tres modelos con los algoritmos seleccionados para compararlos basandonos en sus metricas y escoger el modelo que mejor se adapte a el contexto. Por esta razon, se definieron 3 sets de datos dependiendo de la transformacion en el que se realizo un 70% de entrenamiento y un 30% de los datos para la prueba.

3.1. Modelos Term Frequency Inverse Document Frequency

3.1.1. Regresión logística

Métricas	
Precisión	0.4646
Weighted Recall	0.46
Weighted F1-Score	0.46

El modelo realizado fue una regresión logistica para el set de datos de Term Frequency Inverse Document Frequency. La precisión global del modelo es del 46.46%, lo que significa que el modelo clasifica correctamente el 46.46% de las muestras del conjunto de prueba. El recall es de 0.46 y este indentifica todas las muestras positivas. El F1-Score fue de 0.46 y esta medida combina la precision y el recall, se espera que este valor sea lo más cercano a 1.

3.1.2. Multinomial Naive Bayes

Métricas	
Precisión	0.4582
Weighted Recall	0.46
Weighted F1-Score	0.44

El modelo realizado fue un multinomial Naive Bayes para el set de datos de Term Frequency Inverse Document Frequency. La precisión global del modelo es del 45.82%, lo que significa que el modelo clasifica correctamente el 45.82% de las muestras del conjunto de prueba. El recall es de 0.46 y este indentifica todas las muestras positivas. El F1-Score fue de 0.44 y esta medida combina la precision y el recall, se espera que este valor sea lo más cercano a 1.

3.1.3. Random Forest

Métricas	
Precisión	0.4526
Weighted Recall	0.45
Weighted F1-Score	0.43

El modelo realizado fue un Random Forest para el set de datos de Term Frequency Inverse Document Frequency. La precisión global del modelo es del 45.42% el menor en los resultados del set. El recall es de 0.45 y este indentifica todas las muestras positivas. El F1-Score fue de 0.43 y esta medida combina la precision y el recall. En terminos generales este entre los primeros 3 modelos, este modelo es el que cuenta con peores resultados.

3.2. Count Vectorizer

3.2.1. Regresión Logistica

Métricas	
Precisión	0.4632
Weighted Recall	0.46
Weighted F1-Score	0.46

El modelo realizado fue una regresión logistica para el set de datos de Count Vectorizer. La precisión global del modelo es del 46.32%, lo que significa que el modelo clasifica correctamente el 46.32% de las muestras del conjunto de prueba. El recall es de 0.46 y este indentifica todas las muestras positivas. El F1-Score fue de 0.46 y esta medida combina la precision y el recall, se espera que este valor sea lo más cercano a 1.

3.2.2. Multinomial Naive Bayes

Métricas	
Precisión	0.4650
Weighted Recall	0.46
Weighted F1-Score	0.45

El modelo realizado fue un multinomial Naive Bayes para el set de datos de Count Vectorizer. La precisión global del modelo es del 46.50%, lo que significa que el modelo clasifica correctamente el 46.50% de las muestras del conjunto de prueba. El recall es de 0.46 y este indentifica todas las muestras positivas. El F1-Score fue de 0.45 y esta medida combina la precision y el recall, se espera que este valor sea lo más cercano a 1.

3.2.3. Random Forest

Métricas	
Precisión	0.4333
Weighted Recall	0.43
Weighted F1-Score	0.40

El modelo realizado fue un Random Forest para el set de datos de Count Vectorizer. La precisión global del modelo es del 43.33% el menor en los resultados del set. El recall es de 0.43 y este indentifica todas las muestras positivas. El F1-Score fue de 0.40 y esta medida combina la precision y el recall. En terminos generales este entre los primeros 3 modelos, este modelo es el que cuenta con peores resultados para su set de datos (Count Vectorize).

3.3. Count Vectorizer Dummy

3.3.1. Regresión Logistica

Métricas	
Precisión	0.4594
Weighted Recall	0.46
Weighted F1-Score	0.45

El modelo realizado fue una regresión logistica para el set de datos de Count Vectorizer Binario. La precisión global del modelo es del 45.94%, lo que significa que el modelo clasifica correctamente el 45.94% de las muestras del conjunto de prueba. El recall es de 0.46 y este indentifica todas las muestras positivas. El F1-Score fue de 0.45 y esta medida combina la precision y el recall, se espera que este valor sea lo más cercano a 1.

3.3.2. Multinomial Naive Bayes

Métricas	
Precisión	0.4637
Weighted Recall	0.46
Weighted F1-Score	0.45

El modelo realizado fue un multinomial Naive Bayes para el set de datos de Count Vectorizer Binario. La precisión global del modelo es del 46.37%, lo que significa que el modelo clasifica correctamente el 46.37% de las muestras del conjunto de prueba. El recall es de 0.46 y este indentifica todas las muestras positivas. El F1-Score fue de 0.45 y esta medida combina la precision y el recall, se espera que este valor sea lo más cercano a 1.

3.3.3. Random Forest

Métricas	
Precisión	0.4295
Weighted Recall	0.43
Weighted F1-Score	0.40

El modelo realizado fue un Random Forest para el set de datos de Count Vectorizer Binario. La precisión global del modelo es del 42.95% el menor en los resultados del set. El recall es de 0.43 y este indentifica todas las muestras positivas. El F1-Score fue de 0.40 y esta medida combina la precision y el recall. En terminos generales este entre los primeros 3 modelos, este modelo es el que cuenta con peores resultados para su set de datos (Count Vectorizer Binario).

3.4. Selección del modelo

Basado en las métricas anteriores el modelo que se considera mejor es:

Multinomial Naive Bayes utilizando la transformacion de Count Vectorizer debido a que obtuvo la precisión más alta entre todos los modelos realizados. El Weighted Recall y el Weighted F1-Score, es similar a los modelos que tienen estas métricas más altas.

4. Resultados

Es importante que el algoritmo seleccionado brinda la probabilidad logarítmica de cada palabra en cada clase. Esto quiere decir que las palabras que más influyen en una clase son las que más cercanas a 0 sean. Dado el alto número de palabras en el modelo, se exportaron los resultados en un archivo xlsx y se creó una nueva columna para calcular la diferencia absoluta entre la probabilidad de la clase 1 y la clase 5 y se ordenó de mayor a menor. De esta forma se podía saber cuáles eran las palabras que más influían entre los extremos de las categorías. (Revisar el Anexo 1)

Entre los resultados resaltan algunas palabras como:

- sucio: Con una influencia mayor en la clase 1 que en la clase 5.
- cómoda: Con una influencia mayor en la clase 5 que en la clase 1.
- histórico: Con una influencia mayor en la clase 5 que en la clase 1.
- alfombra: Con una influencia mayor en la clase 1 que en la clase 5.
- moho: Con una influencia mayor en la clase 1 que en la clase 5.
- jardín: Con una influencia mayor en la clase 5 que en la clase 1.
- cucaracha: Con una influencia mayor en la clase 1 que en la clase 5.

De esta forma se obtiene algunos diferenciadores entre las clases como lo puede ser la suciedad, que trae una connotación negativa hacia los huéspedes. La comodidad que es un aspecto importante en el buen servicio para los huéspedes. El hecho que sea histórico tiene influencia en los huéspedes por lo que pueden buscar implementar o mantener este aspecto. El hecho de que no haya alfombras es importante en la percepción de los huéspedes. De igual forma que no haya moho o que no se encuentren cucarachas. El hecho de que exista un jardín es importante para los huéspedes por lo que los hoteles deberían implementar estos espacios.

5. Mapa de actores

El mapa de actores a continuación se construye teniendo en cuenta que el producto creado es para un proyecto se enfoca en el análisis de reseñas de sitios turísticos para determinar la satisfacción de los visitantes.

En cuanto a la organización beneficiada podemos determinar que son el Ministerio de Comercio, Industria y Turismo de Colombia y asociaciones hoteleras como COTELCO y cadenas como Hilton, Hoteles Estelar, y Holiday Inn.

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Gerentes de hoteles y sitios turísticos	Usuario-cliente	Capacidad para identificar áreas de mejora y fortalezas en sus servicios y ofertas.	Dependencia excesiva en el modelo para la toma de decisiones sin considerar factores contextuales o cualitativos no capturados en las reseñas.
Organismos gubernamentales o Empresas privadas	Financiadores	Contribución al desarrollo del turismo y la economía local.	Inversión en un proyecto que no genera los retornos o impactos esperados.
Plataformas de reseñas de turismo y hoteles.	Proveedores de datos	Colaboración en un proyecto que mejora la industria y potencialmente atrae más usuarios a sus plataformas.	Cuestionamientos sobre la privacidad y el uso de los datos de los usuarios.
Turistas	Beneficiarios directos	Mejora en la calidad de los servicios y experiencias turísticas.	Si el modelo no es exacto, podría llevar a recomendaciones que no mejoren su experiencia.
la industria turística	Beneficiarios indirectos	Aumento en la satisfacción del cliente puede traducirse en mejor reputación y más visitantes.	Desalineación entre las mejoras implementadas y las expectativas reales de los visitantes.

6. Trabajo en equipo

Roles y Tareas

- **Líder de Proyecto:** Juan Manuel Pérez

Responsabilidades: Coordinar reuniones, definir plazos, asegurar la distribución equitativa de tareas, y tener la decisión final en caso de desacuerdos. **Tiempo dedicado:** 3 horas. **Retos:** Mantener al equipo en plazo y coordinar la comunicación efectiva entre miembros. **Soluciones:** Establecer un calendario de reuniones regulares y un canal de comunicación constante.

- **Líder de Negocio:** Jeronimo Vargas Rendon

Responsabilidades: Asegurar que el proyecto alinee con los objetivos del negocio, facilitar la comunicación con expertos externos. **Tiempo dedicado:** 10 horas. **Retos:** Sintetizar los requerimientos del negocio en especificaciones claras para el equipo. **Soluciones:** Crear un documento compartido donde se detallen los objetivos y expectativas del negocio.

- **Líder de Datos:** Felipe Núñez

Responsabilidades: Gestionar y disponibilidad los datos, asignar tareas relacionadas con el manejo de datos. **Tiempo dedicado:** 10 horas. **Retos:** Asegurar la calidad y accesibilidad de los datos para todo el equipo. **Soluciones:** Implementar un sistema de almacenamiento en la nube para facilitar el acceso a los datos.

- **Líder de Analítica:** Juan Manuel Pérez

Responsabilidades: Supervisar el desarrollo analítico, garantizar el cumplimiento de estándares y la selección del mejor modelo. **Tiempo dedicado:** 10 horas. **Retos:** Coordinar la experimentación de múltiples modelos y seleccionar el más adecuado. **Soluciones:** Establecer un marco de evaluación de modelos para comparar su rendimiento de manera sistemática.

Reuniones Planificadas

- **Lanzamiento y Planeación:** Definir roles, establecer metodología de trabajo, distribución de tareas y construcción de un calendario de avances de tareas.
- **Ideación:** Discutir los datos explorados, identificar la organización beneficiada y su beneficio por el desarrollo de este proyecto.
- **Seguimiento Semanal:** Revisar avances, resolver dudas, y ajustar el plan según sea necesario.
- **Reunión de Finalización:** Revisar y asegurar que todos los entregables del proyecto estén completos y reflejen el trabajo acordado. Analizar cómo se desarrolló el trabajo en equipo, la gestión del proyecto, y la comunicación entre los miembros. Discutir qué se puede mejorar para la próxima etapa o para futuros proyectos.

Distribución de Puntos:

Suponiendo que todos los miembros han contribuido equitativamente, cada uno recibiría 50 puntos de 150 posibles distribuidos en tres integrantes. Sin embargo, esto puede ajustarse en función de las contribuciones observadas y los retos superados por cada miembro.

Puntos Para Mejorar para la Siguiente Entrega:

Incrementar la frecuencia de las reuniones de seguimiento para monitorear el progreso más de cerca.

Mejorar la documentación del proceso para facilitar la replicabilidad y la comprensión del proyecto.

Anexos:

Anexo 1: Tabla de palabras y sus clases

Palabra/Clase	1	2	3	4	5
sucio	-6.40647851	-6.6554766	-7.44324099	-9.30779976	-10.4383714
pésimo	-6.44786373	-7.5169591	-8.82953536	-10.0009469	-10.4383714
deliciosa	-10.7105436	-9.53186212	-8.72417484	-8.26634588	-6.7877132
peor	-6.43387749	-7.00613347	-7.89126572	-9.30779976	-10.0329063
había	-7.61950115	-7.87363404	-8.08232095	-9.30779976	-11.1315186
espectacular	-10.0173964	-8.43324983	-7.47141187	-7.29289674	-6.57764173
cómoda	-10.7105436	-8.83871494	-8.82953536	-7.51604029	-7.30287723
excelente	-8.22563696	-7.22927703	-6.62004069	-5.71966187	-4.82160034
histórico	-10.7105436	-8.08494314	-7.6594641	-7.12926731	-7.32485613
siquiera	-7.76610463	-7.9737175	-8.72417484	-9.71326486	-11.1315186
alfombra	-7.76610463	-7.82711403	-8.31870973	-9.49012131	-11.1315186
botero	-10.7105436	-9.81954419	-9.92814764	-8.4605019	-7.44263917
grosero	-7.87733026	-8.61557139	-9.23500046	-10.406412	-11.1315186
terrible	-7.30934623	-7.48416928	-7.9822375	-9.15364908	-10.4383714
moho	-8.00249341	-7.9737175	-9.08084978	-11.0995592	-11.1315186
agencia	-8.00249341	-8.35320712	-9.23500046	-10.406412	-11.1315186
ideal	-10.7105436	-9.30871857	-8.46181058	-7.96406501	-7.63501106
desastre	-8.07148628	-8.97224633	-8.82953536	-10.406412	-11.1315186
divertido	-10.7105436	-9.81954419	-8.25417121	-7.51604029	-7.66578272
horrible	-6.99697154	-7.39179596	-8.46181058	-9.30779976	-10.0329063
apreciar	-10.7105436	-8.97224633	-8.62886466	-7.73226339	-7.69753142
cucaracha	-7.45244707	-7.69928066	-9.64046557	-9.49012131	-10.4383714
decepcionante	-8.14559425	-7.74010265	-8.46181058	-9.71326486	-11.1315186
tampoco	-8.14559425	-8.02778472	-7.89126572	-9.49012131	-11.1315186
magnífico	-10.7105436	-10.2250093	-9.64046557	-8.26634588	-7.73032124
obligada	-10.7105436	-9.81954419	-8.31870973	-7.92150539	-7.76422279
jardín	-10.7105436	-8.52026121	-8.13638818	-7.73226339	-7.79931411