Proyecto 1 Analítica de Textos – Turismo de los Alpes

Docente: Haydemar Nuñez

David Cuevas Alba - 202122284
Eduardo José Herrera Alba - 201912865
Samuel Alejandro Jiménez Ramírez - 202116652

Sección: 3

Universidad de los Andes
Departamento de Ingeniería de Sistemas y Computación
2024-10

Tabla de contenido

1. Entendimiento del negocio y enfoque analítico

Objetivos de negocio:

Como equipo, entendimos que el negocio busca obtener información importante para mejorar su servicio a los clientes. El objetivo estaría resuelto si los distintos interesados en el proyecto encuentran un apoyo a la toma de decisiones para cumplir con el requisito de mejorar la calidad del servicio. En general se encuentran dos objetivos concretos:

- Identificar cuáles son los conceptos claves que influyen en la calificación de un usuario para que su reseña sea positiva o negativa, para así tener claras las oportunidades de mejora.
- Obtener un modelo predictivo para obtener la calificación que tendría un lugar usando reseñas que tenga sin calificación o las palabras claves que puedan describir el lugar.

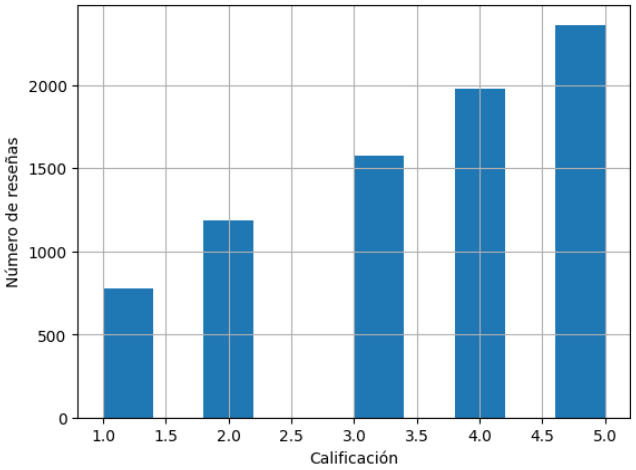
También identificamos que el Ministerio de Comercio, Industria y Turismo de Colombia estaría interesado en el proyecto para generar estrategias que lleven al país a recibir más ingresos por el sector turístico mejorando la calidad de servicio y atrayendo turistas de cualquier parte del mundo, mejorando la calidad de vida de las personas que se dedican a este sector.

Oportunidad/ Problema Negocio	Utilizar un modelo de clasificación de reseñas permite a las empresas comprender mejor las opiniones y necesidades de sus clientes. Esto puede conducir a la identificación de áreas de mejora y la implementación de cambios que aumenten la satisfacción del cliente y fidelicen a los clientes existentes. Esto se hará con ayuda del procesamiento de lenguaje natural para identificar conceptos claves.
Enfoque analítico	Desde el punto de vista del aprendizaje automático, el enfoque analítico para el requerimiento será un modelo de clasificación de reseñas que involucra la recopilación y preprocesamiento de datos, seguido por la extracción de características y la selección del modelo apropiado. Luego, se entrenará el modelo, se evalúa su rendimiento y se ajusta si es necesario.
Rol dentro de la organización	Una entidad del sector turismo se beneficiaría enormemente de un modelo de clasificación de reseñas al poder analizar y comprender de manera rápida y eficiente las opiniones de los clientes sobre sus servicios, destinos turísticos y experiencias.
Contacto con experto externo	En este caso el rol de expertos será desempeñado por los estudiantes de estadística: • Laura Mejía (l.mejiae@uniandes.edu.co) • Laura González (ls.gonzaleza@uniandes.edu.co) La formalidad que definimos fue el de usar un lenguaje muy natural o con las explicaciones técnicas necesarias para que el proyecto sea entendible por cualquier actor. Además de corroborar el enfoque que se le dió al proyecto.

2. Entendimiento y preparación de los datos

En lo que respecta al perfilamiento de los datos a utilizar en el proyecto, se trabajó con un conjunto de reseñas con su correspondiente calificación en un archivo de tipo csv (*comma-separated values*). Para poder entender los datos correctamente se siguieron los siguientes pasos:

1. Se cargaron los datos del archivo en el Notebook utilizando la librería Pandas de Python.
2. Se realizo una exploración inicial de los datos para comprender su estructura. Se encontraron en total 7874 datos en un par reseña calificación.
3. Se crearon estadísticas descriptivas para las clasificaciones, incluyendo la moda, el conteo de palabras y la distribución de clases. A continuación, podemos ver una gráfica de barras que muestra cómo están distribuidas las reseñas en las diferentes categorías.

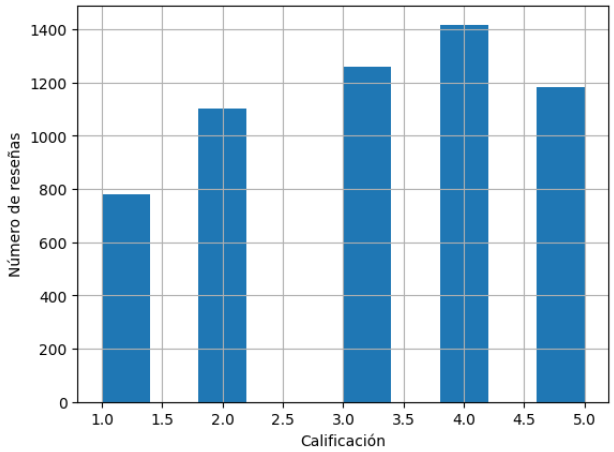


4. Se observó la cantidad de valores faltantes y valores repetidos, igualmente se hizo una búsqueda de reseñas en otro idioma que no fuera español.

Después de este acercamiento a los datos, y observar la calidad de estos se aplicaron múltiples técnicas y tratamientos sobre estos para mejorar el desempeño del modelo, entre estos los siguientes:

1. Remover los valores duplicados o nulos, además de los que no estaban en español para así generar un set de datos para entrenamiento limpio.
2. Remover caracteres que no sean ASCII. Esto se realizó para remover caracteres que los algoritmos no pudieran manejar de manera correcta.
3. Hacer que todas las letras sean minúsculas. Esto se realizó para evitar duplicados de palabras que se diferencien únicamente en mayúsculas. Dando así importancia a solo la palabra como tal y no a cómo esta está escrita.
4. Remover puntuación de las reseñas. Debido a que son caracteres que no nos importan para las tareas que vamos a realizar y pueden dañar el modelo.

5. Reemplazar caracteres numéricos con su versión en palabra. Esto para generar una normalización de los números que se pueden encontrar en las reseñas.
6. Remover las palabras vacías de las reseñas. Estas son palabras que no añaden nada importante a las reseñas y pueden ser conectores de las palabras o muletillas frecuentemente usadas y que no aportan información.
7. Remover caracteres no alfanuméricos. Esto debido a que a nuestros algoritmos no les interesa conocer los caracteres que no son alfanuméricos porque pueden generar errores en el entrenamiento del modelo.
8. Eliminar palabras muy comunes. Decidimos quitar las palabras comunes que hacen parte de las reseñas y se repiten con frecuencia en otras de diferente clase, porque también pueden ser palabras comunes que no aportan información valiosa al calificar una reseña.
9. Se reservaron palabras características de cada clase y se eliminaron de las demás. Esto para solo considerar las palabras que influyen en la calificación de una reseña para considerarlas.
10. Se “nivelo” la cantidad de datos por clase, debido a que la cantidad de datos por clase no era equitativa, lo que sesgaba al modelo. Esto se hizo eliminando filas de las clases con mayor número de ocurrencias donde se eliminaron 2137 reseñas correspondientes a las clases más representadas para nivelar la cantidad de datos. Finalmente, la distribución se muestra en la siguiente gráfica:



Para la limpieza específica de los datos que van a ser pasados al algoritmo realizamos los siguientes pasos:

A. “Tokenizar”(dividir en símbolos/palabras) las reseñas.
B. Se realizó stemming o reducción de palabras a su raíz para aumentar la cantidad de ocurrencias de las palabras.
C. Se “vectorizó” (convertir a un vector/lista de números) las reseñas para pasárselo al modelo de clasificación.

Esto se hizo para pasarle al algoritmo datos que pudiera entender, ya que por su naturaleza solo recibe datos numéricos. Además, se dividieron los datos en entrenamiento y test, para evaluar el modelo con datos que no conociera antes para evaluar el sobreajuste.

3. Modelado y evaluación

Después de realizar el procesamiento de lenguaje natural pasando por la limpieza y preparación llegó la hora de realizar modelos de aprendizaje automático. Para ello, usamos la tarea de clasificación y elegimos los siguientes 3 algoritmos:

- Regresión Logística realizada por Samuel Jiménez:

La regresión logística es un algoritmo de clasificación que tiene dos tipos, el de clases binomiales y multinomiales. Para nuestro caso utilizamos clasificación multinomial. La regresión logística se usa para predecir la probabilidad de que un dato pertenezca a cierta categoría o clase.

Para predecir a qué clase pertenece un dato, el algoritmo realiza la probabilidad para cada una de las clases y seleccionando la que tenga una probabilidad más alta.

El algoritmo puede recibir datos dispersos o densos. Como parámetros del algoritmo modificamos 3 parámetros de los por defecto los cuales son: *penalty*, utilizado para evitar el sobreajuste escogiendo el tipo de penalización; *solver*, que indicará el algoritmo de optimización utilizado; y finalmente C que es el inverso de la fuerza de regularización donde números más pequeños indican una mayor regularización.

En la búsqueda de parámetros encontramos que los que mejor se acoplaban a nuestros datos eran: un C de 1, un *penalty* l2 y un *solver* saga.

- Gaussian Naive Bayes realizada por David Cuevas:

El algoritmo de Gaussian Naive Bayes es una variante del algoritmo de Naive Bayes que es utilizado cuando las características se distribuyen de manera normal. Se usa una estimación de la media y la desviación estándar de cada característica en cada clase para calcular las probabilidades condicionales necesarias para la clasificación.

Para predecir la clase a la que pertenece cierto dato, el modelo luego del entrenamiento utiliza la probabilidad de que pertenezca a cada clase utilizando el teorema de Bayes. Finalmente se utiliza el principio de máxima verosimilitud para determinar la clase más probable de la nueva muestra.

Se modificó el parámetro de var_*smoothing* por 8 puesto que fue el que mejor se adaptó al modelo y nuestros datos, este parámetro es un factor de suavizado para evitar problemas de división por cero al calcular varianzas.

- Ridge Classifier realizada por Eduardo Alba:

El algoritmo de Ridge *Classifier* es un clasificador lineal regularizado que es utilizado con la regresión ridge para entrenarse. En este algoritmo se usa la penalización de regularización tipo L2 para evitar el sobreajuste al imponer una restricción sobre su magnitud.

Este algoritmo para realizar la predicción primero hace un ajuste de coeficientes, luego utiliza estos mismos para calcular la puntuación para cada clase y finalmente selecciona la clase con la puntuación más alta.

Para el uso de este algoritmo modificamos los parámetros: *alpha*, que controla la fuerza de regularización; tol, que define la convergencia del algoritmo de optimización; y el *solver*, definido como *sparse_cg* en nuestro caso, que se usa para referiste al método de gradiente utilizado para resolver el problema de optimización.

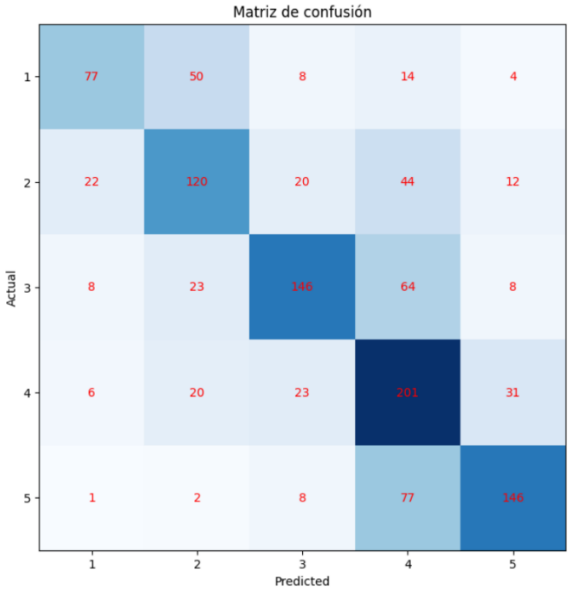
Métricas obtenidas para cada uno de los tres algoritmos:

	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Gaussian Naive Bayes	73%	57%	67%	52%	66%	50%
Ridge Classifier	83%	60%	84%	63%	83%	61%
Regresión Logísitca	88%	61%	88%	63%	88%	61%
	Exactitud		**Precisión**		**F1 Score**	

Estas fueron las métricas que obtuvimos al evaluar los distintos algoritmos implementados, vamos a dar una pequeña definición de cada una de las métricas:

- **Exactitud:** Conocida en inglés como "accuracy", esta métrica se calcula dividiendo el número de predicciones acertadas por el total de predicciones hechas. El resultado es el porcentaje que representa cuán a menudo el modelo realiza predicciones correctas.
- **Precisión:** Esta métrica evalúa qué tan precisas son las predicciones positivas del modelo para cada clase. Se determina por la proporción de muestras correctamente identificadas como pertenecientes a una clase, dividida entre todas las muestras que el modelo ha clasificado en esa clase.
- **F1 Score:** El F1 score es un indicador que sintetiza la precisión y el recall de un modelo para una clase específica. El recall mide la habilidad del modelo para detectar todas las instancias reales de una clase. El F1 score es crucial porque refleja el equilibrio entre estas dos métricas, ofreciendo una visión comprensiva de la performance del modelo.

Como resultado, obtenemos que el modelo de Regresión Logística es el más adecuado a la hora de predecir los valores necesarios puesto que sus métricas son más altas en comparación con los otros modelos, en la siguiente gráfica podemos ver la matriz de confusión en la que evidenciamos que tiene un buen manejo de errores.



El manejo de errores no genera discrepancias tan grandes puesto que se puede evidenciar que las predicciones erróneas se agrupan sobre la predicción esperada, es decir que para una predicción esperada de 3 la mayoría de los errores están sobre 2 y 4, dando así una diferencia pequeña en comparación.

4. Resultados

Con base en el paso anterior, evidenciamos que el modelo de regresión logística obtuvo mejores métricas. Este modelo tiene un desempeño aceptable para asignar cada clase apropiadamente o en un valor muy cercano, por lo que puede usarse.

Esto permitirá a las diferentes organizaciones realizar predicciones del desempeño de un establecimiento dado, con base a sus características descritas en una reseña.

A continuación, vamos a evaluar la nube de palabras de cada una de las categorías de las reseñas.

Nube de palabras para las reseñas con 1 estrella:

Las palabras más recurrentes en las reseñas de 1 estrella son negativas en su mayoría, podemos evidenciar que adjetivos como sucio, tarde o pésimo son recurrentes. Al igual que elementos como aire acondicionado, toallas, mesa y piso también. Podríamos sugerir que una de las estrategias a implementar sería la de mantener los espacios limpios, ser puntuales con la hora de entrega del alojamiento, además de tener un buen inventario para generar experiencias agradables y mejorar su servicio.

Nube de palabras para las reseñas con 2 estrella:

Para las reseñas de 2 estrellas acaba siendo más importante términos relacionados al precio y a los pagos, además de cuestiones de tiempo y detalles que arruinaron la experiencia. Igualmente sugerimos tener puntualidad en los servicios para evitar este tipo de inconvenientes. Adicionalmente, como vemos tantas palabras relacionadas con el precio y sus métodos de pago sugerimos manejar precios estándar o acordes con la calidad de servicio, además de tener distintos métodos de pago para facilitar esta experiencia.

Nube de palabras para las reseñas con 3 estrellas:

Para las reseñas de 3 estrellas se encuentran palabras positivas, pero no al extremo (como excelente, perfecto, ideal, etc), lo cual tiene sentido, ya que se representan experiencias aceptables, más no excepcionales. Siguen apareciendo palabras que expresan temporalidad y pagos así que haríamos las mismas recomendaciones. Adicionalmente, aparece el adjetivo de vieja, con lo que sugeriríamos tener un inmobiliario nuevo o por lo menos con un mantenimiento frecuente para no notar sus años.

Nube de palabras para las reseñas con 4 estrellas:

En el caso de las reseñas con 4 estrellas las palabras están relacionadas a detalles de satisfacción, como “vale” y “pena” (que deben venir de vale la pena), donde se ve que son usuarios satisfechos con su experiencia. También se evidencian palabras como grandes, buffet y limpio por lo que recomendaríamos tener alojamientos de gran tamaño y ofrecer alternativas alimenticias como *buffets*, además de siempre procurar buenas limpiezas.

Nube de palabras para las reseñas con 5 estrellas:

Para 5 estrellas hay una gran selección de palabras positivas al extremo, que demuestran la satisfacción de los usuarios con la experiencia, al punto de usarlas para describir su estadía. También vemos palabras como amables, limpias, atentas y sabor con las que recomendamos a los lugares turísticos tener buena atención al cliente para mejorar sus calificaciones, además de comida agradable y lugares siempre limpios.

En cuanto a las métricas de calidad del mejor modelo podemos evidenciar que con las métricas obtenidas es de esperar un rendimiento al rededor del 60%, esto quiere decir que nuestro modelo mayormente es correcto a la hora de evaluar una reseña u opinión sobre un lugar.

Esta información es valiosa para el negocio ya que les informa sobre los conceptos clave de las calificaciones y así interesarse en mejorar para satisfacer a sus clientes.

Por parte de la clasificación, esta información es valiosa para obtener métricas de rendimiento y calidad de los servicios, además de mediante recomendaciones sin valor numérico pueden obtenerlas sin interpretarlas humanas.

5. Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Empresas de turismo	Financiadores-Cliente-Usuario	Mecanismo de toma de decisiones basado en métricas obtenidas estudiando a los usuarios del servicio. También podrán encontrar cuáles son las principales características que hacen un servicio bueno para mejorar su calidad.	Si el modelo no funciona es dinero mal invertido y pudo dejarse de hacer un proyecto con mayor impacto y viabilidad. Además, puede generarse una decisión incorrecta debido a un modelo mal implementado.
Equipo de desarrollo	Proveedores	Se asegura que los estándares de calidad de los productos desarrollados se cumplan, que abarca aspectos como la seguridad y privacidad de los datos utilizados.	Manejo incorrecto de los datos que lleve a la violación de la privacidad de los datos
Viajero	Beneficiados	Puede obtener un mejor servicio en sus viajes y tener una satisfacción más como turista porque las empresas buscan mejorar su calidad. Además, sus valoraciones críticas tendrán valor para los hoteles, según se considerarán.	Puede que el modelo sugiera erróneamente que cambios hacer a los servicios y empeore la experiencia, en un caso no tan catastrófico podría no notar interés en el turismo porque no mejoran sus falencias y dejar de usar estos servicios.
Ministerio de Comercio, Industria y Turismo de Colombia	Financiador - Cliente	Puede generar campañas para que los lugares turísticos implementen estrategias que ayuden a mejorar el servicio. Con esto obtendrían un incremento en los ingresos del sector por medio del aumento de turistas y visitantes al país.	Podrían financiar con dinero público estrategias no eficientes que no generen la retribución de objetivos deseada. Igualmente, la toma de decisiones basa en productos mal diseñados puede llevar a decisiones equivocadas y poco efectivas.

6. Trabajo en equipo

Trabajamos con los siguientes roles en el proyecto para mejorar nuestro desempeño

- Líder de analítica: Eduardo Alba – Algoritmo: *Ridge Classifier* – 34 Puntos. 11 horas dedicadas.
    - Planteamiento inicial de limpieza y entendimiento de datos
    - Integración de notebooks individuales
    - Investigación de librerías principales y selección de librerías auxiliares
    - Descripción general de pasos en el notebook
    - Obtención de métricas finales y el mejor modelo
    - Predicción de datos no etiquetados
- Líder de negocio: Samuel Jiménez – Algoritmo: Regresión Logística – 33 Puntos. 10 horas trabajadas.
    - Planteamiento de objetivos de negocio
    - Análisis de modelos obtenidos y sus métricas
    - Análisis de resultados.
    - Adicionar información para el negocio en el mapa de actores y la limpieza y entendimiento de datos.
    - Sugerencia de estrategias a ser implementadas por los sitios turísticos.
- Líder de datos: David Cuevas Alba – Algoritmo: *Gaussian Naive Bayes* – 33 Puntos. 12 horas dedicadas.
    - Selección de palabras descriptivas
    - Redistribución de los datos
    - Pruebas preliminares de la distribución con otros algoritmos para evaluar desempeño de los datos
    - Descripción de la limpieza y entendimiento de los datos
    - Descripción de los resultados de los modelos creados
    - Mapa de actores

Repartimos los 100 puntos equitativamente, ya que consideramos que trabajamos a la par y todos respondimos por la entrega final, por lo que cada uno obtuvo 33 puntos.

Se realizaron las siguientes reuniones:

1. Reunión de lanzamiento y planeación: Para definir roles y forma de trabajo del grupo.
2. 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

Referencias

GeeksforGeeks. (s/f). GeeksforGeeks. Recuperado el 7 de abril de 2024, de https://www.geeksforgeeks.org/

Scikit-learn. (s/f). Scikit-learn.org. Recuperado el 7 de abril de 2024, de https://scikit-learn.org/stable/