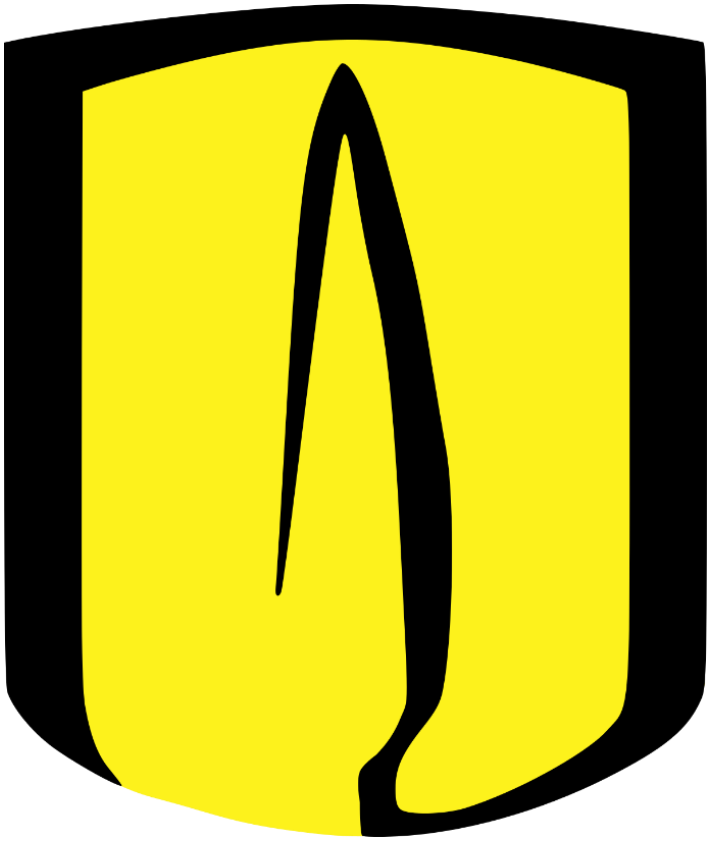
Etapa 1 - Construcción de modelos de analítica de textos

Turismo de los Alpes



Integrantes:

María Camila Luna Velasco - 201920993

Juan Manuel Jauregui Rozo - 201922481

Ana Sofía Castellanos Mosquera - 202114167

Universidad de Los Andes

2024-2

Tabla de contenido

1. Entendimiento del negocio y enfoque analítico 3
1.1 Cronograma propuesto: 5
2. Entendimiento y preparación de los datos 5
3. Modelado y evaluación 7
4. Resultados 10
4.1 Análisis Cuantitativo 10
4.2 Análisis Cualitativo 10
5. Mapa de actores relacionado con el producto de datos 11
6. Trabajo en equipo 12
7. Entregables 13
8. Referencias 13

1. Entendimiento del negocio y enfoque analítico

A continuación, se presenta una tabla que contiene la definición de los objetivos y los criterios de éxito desde el punto de vista del negocio. Adicionalmente, se presenta el impacto que tiene este proyecto en Colombia para uno de los actores del sector turismo. Por otro lado, se describe el enfoque analítico para alcanzar los objetivos del negocio. Finalmente, se presenta el análisis del proyecto realizado en conjunto con los estudiantes de estadística asignados.

Oportunidad/Problema de negocio	La organización en este caso el Ministerio de Comercio, Industria y Turismo de Colombia busca un mecanismo para generar estrategias de mejora y aumentar la popularidad de los sitios a partir de la calificación del sitio turístico.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar	Desde una perspectiva de aprendizaje automático el requerimiento de la organización se puede traducir como una tarea de clasificación de textos, dado que se brinda un conjunto de textos de reseñas que están acompañados por una clase entre el 1 y 5, asociados a la calificación dada a los sitios turísticos. Cuando llega una reseña nueva, como se menciona en el enunciado, la organización quiere tener un mecanismo para saber la calificación que tendrá un sitio, que se puede ver como el promedio de las calificaciones de las reseñas del lugar. Para cumplir con la tarea, en primer lugar, se propone un preprocesamiento de los datos para lo cual se eliminan duplicados, palabras conectoras, números y caracteres no ascii a la vez que se propone un lematizador para obtener mejores resultados. Tras esto, se proponen tres algoritmos de procesamiento de lenguaje natural los cuales son: Support Vector Machines, Regresión Logística y Decision Trees. Para la aplicación de los algoritmos se propone la partición del conjunto de train en dos uno de train y otro de prueba. Así mismo se buscan hiperparámetros mediante K-fold cross validation y grid search. Finalmente, a partir del mejor modelo, se genera la predicción de los datos
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Este proyecto podría tener un impacto significativo en Colombia en varios aspectos: **Turismo sostenible:** Al analizar las características de los sitios turísticos que atraen a los visitantes y aquellos que no lo hacen, se pueden identificar oportunidades para promover un turismo más sostenible, enfocado en la conservación del medio ambiente y la cultura local.

	Desarrollo económico: Al mejorar la popularidad y la calidad de los sitios turísticos, se puede aumentar el flujo de turistas, lo que a su vez generaría ingresos adicionales para las comunidades locales, empresas de turismo y el país en general. **Empleo:** Un aumento en el turismo podría generar nuevas oportunidades de empleo en sectores como la hospitalidad, la gastronomía, el transporte y las artesanías, beneficiando a las comunidades locales y contribuyendo al crecimiento económico. **Promoción de la cultura y el patrimonio:** Al identificar las características que hacen atractivos a ciertos sitios turísticos, se puede destacar y promover la riqueza cultural, histórica y natural de Colombia, lo que contribuiría a preservar y valorar el patrimonio del país. **Desarrollo regional:** El análisis de los sitios turísticos en diferentes municipios de Colombia podría ayudar a equilibrar el desarrollo económico y turístico en todo el país, promoviendo el turismo en regiones menos visitadas y descentralizando la actividad económica. En resumen, este proyecto podría tener un impacto positivo en Colombia al mejorar la competitividad turística, generar empleo, promover la conservación del medio ambiente y la cultura, y contribuir al desarrollo económico y regional del país.
Contacto con experto externo al proyecto y detalles de la planeación	**Entender las métricas de evaluación:** Familiarizarse con las métricas de evaluación utilizadas, como precisión, recall, F1-score, y área bajo la curva ROC, para comprender cómo se desempeña el modelo en términos de diferentes aspectos como la exactitud, y el equilibrio entre precisión y recall. **Contextualizar los resultados:** Tener en cuenta el contexto del problema que se está abordando y considerar si las métricas de evaluación obtenidas son satisfactorias en ese contexto específico. **Analizar las predicciones erróneas:** Examinar las predicciones erróneas realizadas por tu modelo para identificar patrones comunes o características específicas que podrían estar contribuyendo a errores sistemáticos. Esto puede proporcionar insights valiosos sobre áreas donde tu modelo podría mejorar. **Visualizar los resultados:** Utiliza visualizaciones como matrices de confusión, curvas ROC, gráficos de precisión-recall y diagramas de dispersión para visualizar y comprender mejor el rendimiento del modelo y su capacidad para separar las diferentes clases **Validar la interpretación**: Si el modelo es lo suficientemente complejo como para que su interpretación no sea inmediata, considerar métodos para explicar como la generación de reglas de decisión o el uso de técnicas

	de interpretación de modelos para comprender mejor cómo se están tomando las decisiones. **Iterar y mejorar**: Utilizar los insights obtenidos de la interpretación de los resultados para iterar sobre el modelo, ajustar hiperparámetros y mejorar su rendimiento continuamente. El proceso de interpretación y mejora iterativa es fundamental para desarrollar modelos de alta calidad y robustos. En cuanto a la planeación se propone una fecha de reunión para

1.1 Cronograma propuesto:

Este es el cronograma que se siguió para poder completar la totalidad del proyecto en los plazos esperados.

Fecha	Actividad
1/04/2024 - 12:30m - 1:30pm	Reunión de lanzamiento y planeación
2/04/2024 - 7:00pm - 8:00pm	Reunión de ideación
3/04/2024 - 6:30pm - 7:30pm	Reunión de seguimiento
5/04/2024 – 6:00pm – 7:00pm	Reunión de finalización
6/04/2024 – 10:00am – 11:00am	Reunión con estadística

2. Entendimiento y preparación de los datos

A partir del conjunto de entrenamiento del tipo 2 se realiza un perfilamiento y análisis de calidad de los datos. A partir de este análisis se encuentran estadísticas e información relevante entorno al conjunto de datos. En primer lugar, se encuentra que el tamaño del dataset es de 7875 columnas con dos columnas, una correspondiente a la reseña y otra con la clase anotada.

Como primer acercamiento al perfilamiento de los textos, se buscan las palabras más frecuentes que aparecen en la reseña, dado que a este punto no se ha aplicado ninguna técnica de procesamiento y se tienen los datos en crudo, se observa que las palabras más frecuentes son usualmente conectores tales como ‘de’, ‘y’, ‘la’, ‘el. ‘en’, ‘Es’, entre otros (ver nube de palabras naranja en tablero control 1). Esta información no es relevante al momento de hacer una clasificación y es por ello por lo que es fundamental aplicar la eliminación de stopwords.

Tras este análisis se realiza un pandas profiling, a través del cual se observa que hay datos duplicados, por lo que este paso se incluye en la etapa de procesamiento. De igual manera, como se observa en el tablero de control, el tamaño de las reseñas puede alcanzar los 10.000 caracteres.

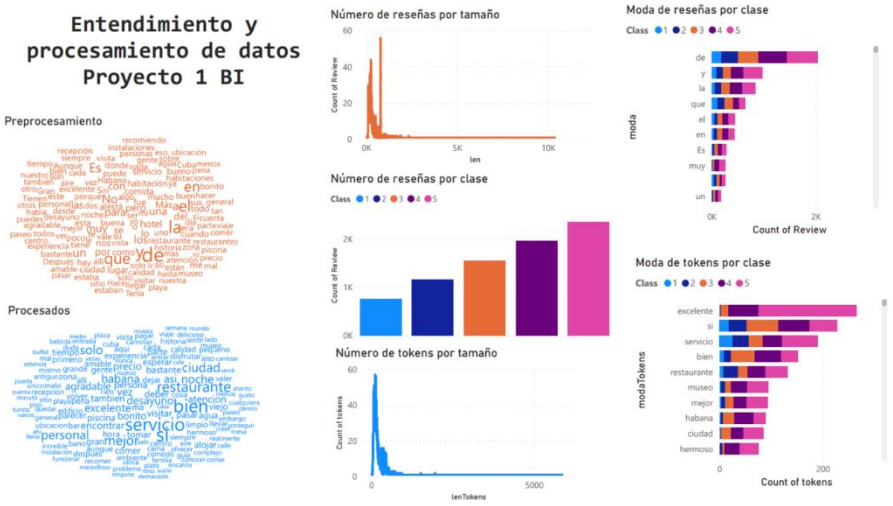
Tras el análisis de los datos se procede a realizar el proceso de preparación y transformación de datos, el cual consta de:

1. Eliminar duplicados y números
2. Aplicar procesamiento pequeño: Eliminación de mayúsculas y no ASCII

3. Lematización
4. Tokenización a los lemas del punto 3
5. Aplicar preprocesamiento completo a los tokens del punto 4. Esto es eliminar puntuación, eliminar palabras de parada (en español) y dejar todo en minúscula (en caso de que el lemmatizador haya generado palabras en mayúscula).

Se aplica un lematizador con el objetivo de transformar los textos en su lema y evitar tener palabras que tienen el mismo significado pero que cuentan con formas gramaticales distintas. El grupo se decantó por la opción del lemmatizador ya que esto permite reducir la cantidad de palabras del vocabulario a la vez que no se pierde el significado de la palabra, caso contrario a lo que sucede al aplicar un stemming donde se obtiene la raíz y se puede perder la significación.

Por otro lado, es importante mencionar que el lematizador se aplica antes del preprocesamiento que elimina palabras conectoras y puntuación, ya que el rendimiento tras aplicar los algoritmos es ligeramente mayor; esto se debe a que Stanza, la librería usada, requiere de las palabras conectoras para comprender las palabras a lematizar.



Tablero de control 1

Tras el procesamiento de los datos estos se vuelven a perfilar a través de los gráficos mostrados en el notebook y en el tablero de control 1. Al respecto se observa que cambia drásticamente la nube de palabras de forma positiva, puesto que no surgen palabras conectoras, sino palabras con un significado positivo/ negativo, que permitirán obtener mejores resultados en el modelo. Así mismo, el tamaño de las reseñas una vez procesadas se reduce de un máximo de 10.000 caracteres a 5000.

Finalmente, con la gráfica 'Numero de reseñas por clase' del tablero de control, se observa que el dataset está desbalanceado máximo en una relación 1:3. Con el objetivo de no perder información que puede ser relevante se decide dejar los datos con el desbalance.

Se hicieron tres análisis: aplicando lemas previos al procesamiento fuerte (paso 4), aplicando lemas después de este procesamiento y sin aplicar lemas. El análisis por

aparte se encuentra en la carpeta AlgoritmosDetallados “Entendimiento de datos”, “Entendimiento de datosLemmaFirst” y “Entendimiento de datosSinLemas”.

3. Modelado y evaluación

Tras el entendimiento y preparación de los datos se proponen tres algoritmos diferentes para solucionar el problema, los cuales se describen detalladamente a continuación junto con los resultados obtenidos con cada uno. Adicionalmente, para cada algoritmo se probaron distintos métodos para preparar los datos incluyendo: Vectorización por conteo, vectorización TFIDF, además, se hizo el uso de lemas de tres formas (antes del preprocesamiento, después y sin lemas). A continuación, se muestran los mejores resultados obtenidos en cada algoritmo, así como du descripción:

Regresión Logística: Juan Manuel Jáuregui Rozo

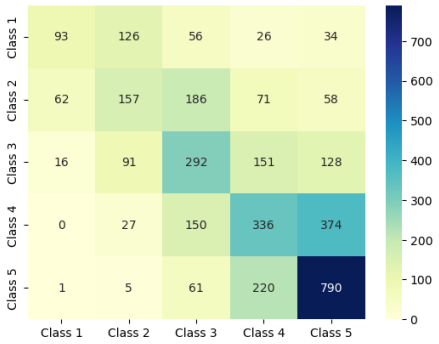
Descripción del Algoritmo: La regresión logística es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación binaria o clasificación multiclase. Este algoritmo utiliza una función llamada “función sigmoide” que modela la probabilidad de que una observación pertenezca a una clase particular. Este algoritmo se entrena ajustando los coeficientes del modelo para minimizar las diferencias entre las probabilidades predichas y las etiquetas verdaderas. En este caso, el algoritmo fue utilizado sin parámetros y con parámetros para ver en qué caso se obtenían mejores resultados. Los parámetros utilizados fueron *C* que es un *float* asociado al inverso de la fuerza de regularización y *solver* que es un algoritmo para usar en el problema de optimización.

Justificación: Este algoritmo se utilizó ya que es una buena elección para realizar una clasificación al igual que el algoritmo SVM. Adicionalmente, este algoritmo es bueno en términos de eficiencia computacional, robustez ante valores atípicos e interpretación de coeficientes. En ese sentido, dado el contexto del proyecto, este algoritmo resulta ser una gran opción para predecir las clasificaciones de las reseñas proporcionadas.

Resultados obtenidos: Para llegar a los resultados obtenidos se hizo búsqueda de hiperparámetros, y se obtuvo el mejor resultado con vectorización tipo TFIDF y aplicando los lemas antes del preprocesamiento de datos:

Exactitud: 0.48. Esto significa que alrededor del 48% de las clasificaciones del modelo son correctas.
Recall: 0.42. Indica que el modelo es capaz de recuperar correctamente el 42% de todas las instancias positivas en los datos.
Precisión: 0.46. Mide la proporción de verdaderos positivos entre todas las instancias que el modelo clasifica como positivas, es del 46%.
Puntaje F1: 0.43. Medida de la precisión y el *recall* son relativamente balanceados, es de alrededor del 43%.

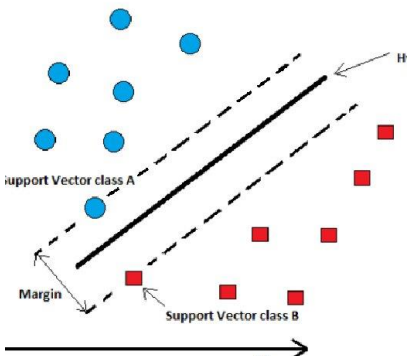
A continuación, se puede observar la matriz de confusión generada para las métricas enunciadas anteriormente.



Como se puede ver, la Clase 5 es la que realiza la mejor clasificación y la peor clasificación es la de la Clase 1. Sin embargo, el puntaje F1 no es tan alto como el de otros algoritmos probados.

Support Vector Machines: Ana Sofía Castellanos Mosquera

Descripción del Algoritmo: El algoritmo SVM es un algoritmo de machine learning supervisado, que permite realizar la clasificación de datos en diferentes clases. El objetivo general del algoritmo es reducir el margen (ver imagen 1), definido como la distancia entre uno (o varios) punto(s) representativo(s) de cada clase y la curva (línea) que clasifica a la clase de las demás. El punto representativo de cada clase se conoce como el Support vector y se define como el punto más cercano a la línea(s) trazada.



Algoritmo SVM

Por defecto el algoritmo hace uso de una clasificación lineal, sin embargo, para datos que tienen clases separadas de forma no lineal se puede hacer uso del kernel para realizar separación de clases de forma que se considere funciones polinomiales, que mida la similaridad como una función exponencial decayente (rbf) o como una sigmoide.

Justificación: Se elige el algoritmo SVM ya que este cumple con la tarea de clasificación supervisada determinada dentro del enfoque analítico. De igual manera, algunas publicaciones científicas como la de Thorsten Joachims afirman que los métodos de clasificación SVM son bastante robustos, presentan una ganancia en cuanto a rendimiento y pueden ser usados sin la necesidad de configuración extensiva de parámetros (grid search).

Resultados obtenidos: El algoritmo se ejecuta de 3 formas diferentes. En primer lugar, se ejecuta con parámetros aleatorios y suponiendo una clasificación lineal. En segundo lugar, se ejecuta con una búsqueda de hiperparámetros simple a partir de la partición x_val y y_val. En tercer lugar, se ejecuta con búsqueda de hiperparámetros mediante K cross-fold validation a tres particiones.

Con la primera forma se obtiene una exactitud del 48%, una precisión del 45%, un recall del 43% y una puntuación f1 del 44%. Si bien los resultados no superan el 50%, es importante resaltar que en general los modelos dada la distribución del conjunto

de datos no logran tener un resultado tan alto. De igual manera puede que los datos estén mal anotados lo que impide una buena clasificación.

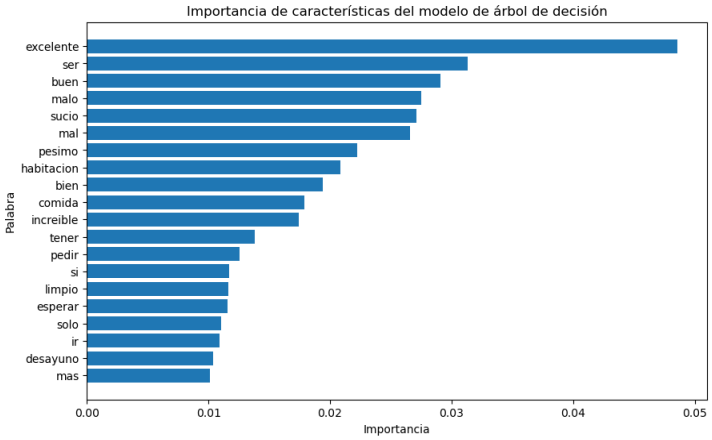
Con la segunda forma, el método falla rotundamente, ya que no realiza ninguna clasificación de las clases 1 ni 2, por lo que el rendimiento baja bastante. Debido a esto se implementa el tercer método con el que se encuentren los mejores resultados con una exactitud del 47%, una precisión del 46%, un recall del 42% y una puntuación f1 del 43%.

A partir de estos resultados y dado que este modelo resulta ser el mejor se pueden obtener las características (palabras) más relevantes por categoría. De esta manera, se observan que la clase 1 corresponde a una pésima calificación que sube hasta la clase 5 que es una calificación excelente (ver sección de resultados).

***Decision Trees*: María Camila Luna Velasco**

Descripción del Algoritmo: Los árboles de decisión son un algoritmo de aprendizaje supervisado utilizado principalmente para tareas de clasificación, aunque también se pueden utilizar para problemas de regresión. Se basan en la idea de dividir repetidamente el conjunto de datos en subconjuntos más pequeños, utilizando reglas de decisión basadas en características, con el objetivo de clasificar correctamente las instancias de datos. Se decidió usar este, porque los árboles de decisión son modelos fácilmente interpretables, lo que significa que puedes entender cómo se toman las decisiones en función de los atributos de las reseñas. Esto nos permite identificar directamente qué aspectos específicos son más importantes para los clientes turísticos e identificar los factores más relevantes que influyen en la satisfacción del cliente en la industria turística. Esto te permite enfocar tus esfuerzos en áreas específicas que tienen un mayor impacto en la experiencia del cliente.

Resultados obtenidos: Para llegar a los resultados obtenidos se hizo búsqueda de hiperparámetros, y se obtuvo el mejor resultado con vectorización tipo TFIDF y aplicando los lemas antes del preprocesamiento de datos:



Exactitud: 0.38. Esto significa que alrededor del 38% de las clasificaciones del modelo son correctas.
Recall: 0.38. Indica que el modelo es capaz de recuperar correctamente el 38% de todas las instancias positivas en los datos.
Precisión: 0.38. Mide la proporción de verdaderos positivos entre todas las instancias que el modelo clasifica como positivas, es del 38%.
Puntaje F1: 0.36. Medida de la precisión y el *recall* son relativamente balanceados, es de alrededor del 36%.

En cuanto a las palabras más relevantes según el modelo. Se obtuvo que para los turistas es muy importante la habitación que les toque, la comida que sea ofrecida y el desayuno en específico. Además, son relevantes los tiempos de espera y que los lugares donde estén sean aseados. Hay palabras como “Malo” que se encuentran en el top 10, lo que indica que debe haber una mejora general del servicio.

4. Resultados

A partir del desarrollo de los tres algoritmos y de los resultados presentados en la sección anterior, es claro que el mejor algoritmo es *Supporting Vector Machines.* A continuación, se explicarán las razones detrás de esta afirmación.

4.1 Análisis Cuantitativo

Analizar las métricas que arroja el algoritmo es fundamental para entender los resultados del modelo aplicado, es decir, la efectividad en la tarea de categorizar las reseñas de los sitios turísticos. Las métricas presentadas anteriormente, las cuales son exactitud, recall, precisión y puntaje F1 dan información sobre las predicciones que hace el modelo sobre las reseñas y como se puede comparar eso con las etiquetas del conjunto de datos. Las etiquetas son las categorías reales a las que pertenece cada una de las reseñas. Ahora bien, se puede realizar un análisis de cada una de las métricas empezando por exactitud, pasando al recall, siguiendo con la precisión y terminando con el puntaje F1. Primero, el modelo clasificó correctamente aproximadamente la mitad de las reseñas. Esto significa que el modelo logró acertar el 47% de las veces si una reseña es positiva o negativa (teniendo en cuenta las diferentes clases definidas). Segundo, el modelo logró identificar correctamente el 44% de las reseñas positivas entre todas las reseñas positivas del conjunto de datos. Tercero, de todas las reseñas clasificadas como positivas, el modelo acertó un 46% de las veces. Finalmente, el puntaje F1 fue de 44.5% lo cual hace referencia al balance entre el recall y la precisión. Esto sugiere que el modelo tiene un equilibrio al momento de identificar correctamente las reseñas positivas y negativas.

A continuación, se pueden observar las palabras que más se repiten dentro del conjunto de datos. En este caso, estas palabras fueron las que más influyeron para calificar las diferentes reseñas como positivas o negativas.

4.2 Análisis Cualitativo

A partir de las métricas obtenidas y de estas palabras clave la organización puede desarrollar varias estrategias. Primero, se pueden identificar áreas de mejora. A partir de las palabras clave que están en las reseñas clasificadas como negativas, se pueden mejorar áreas, por ejemplo, de limpieza o calidad del servicio. Segundo, se pueden desarrollar estrategias de marketing que promuevan los aspectos positivos encontrados en las reseñas clasificadas como positivas correctamente. De esta manera, se pueden atraer más turistas y mejorar las ganancias de la organización. Finalmente, se pueden desarrollar estrategias que se encarguen de mejorar la implementación de este tipo de proyectos para caracterizar cada vez mejor a los turistas y entender que les gusta, que les parece bueno y como complacerlos.

El video que tiene más información sobre los resultados se encuentra disponible aquí: https://youtu.be/4OECNeO6JBw

5. Mapa de actores relacionado con el producto de datos

Se eligen a los hoteles en general como organización benefactora del producto de datos y se identifican los siguientes roles dentro de la organización:

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Gerente de Marketing	Cliente Financiador	Utilización de análisis para identificar destinos turísticos populares y diseñar estrategias de marketing dirigidas a promover esos destinos, lo que puede aumentar el número de turistas y generar mayores ingresos para el hotel.	Riesgo de malinterpretación de los datos o conclusiones incorrectas, lo que podría llevar a la implementación de estrategias de marketing ineficaces o contraproducentes.
Director de Operaciones	Beneficiario	Mejora en la gestión de recursos y personal en el hotel, gracias a la identificación de áreas de oportunidad y puntos débiles detectados por el modelo analítico. Esto puede conducir a una experiencia de turismo más satisfactoria para los huéspedes y una mayor eficiencia operativa.	Riesgo de resistencia al cambio por parte del personal o dificultades en la implementación de nuevas políticas y procesos en el hotel.
Personal de Recepción	Beneficiario	Mejor comprensión de las preferencias y necesidades de los huéspedes a través del análisis de datos, lo que les permite ofrecer un servicio más personalizado y satisfactorio.	Riesgo de sobrecarga de información o dificultades en la implementación de recomendaciones del análisis de datos debido a limitaciones de tiempo o recursos.

Equipo de Limpieza	Beneficiario	Identificación de áreas críticas que requieren atención y mantenimiento, lo que puede ayudar a mantener un ambiente limpio y cómodo para los huéspedes.	Riesgo de falta de recursos o tiempo para abordar todas las áreas identificadas como problemáticas, lo que podría afectar la calidad del servicio ofrecido.
Chef y Equipo de Cocina	Beneficiario	Utilización de datos sobre las preferencias gastronómicas de los huéspedes para adaptar el menú y ofrecer opciones que satisfagan sus necesidades y deseos.	Riesgo de dificultades en la implementación de cambios en el menú o en la gestión de inventario debido a restricciones presupuestarias o logísticas.

6. Trabajo en equipo

Los roles asignados a cada uno de los miembros, así como las tareas que hizo cada uno, se muestran a continuación.

María Camila Luna Velasco (33 puntos)

Rol: Líder de analítica

Tareas: Se encargó de revisar las alternativas posibles y algoritmos que podrían dar mejores resultados. También se encargó de crear y dirigir la implementación de estrategias analíticas para abordar los objetivos comerciales y las necesidades de la organización Turismo de los Alpes. Otra tarea fue identificar oportunidades para mejorar los procesos analíticos existentes y desarrollar e implementar soluciones innovadoras para aumentar la eficiencia y la precisión. Así mismo, se encargó de guiar al equipo en cuanto a técnicas estadísticas y de aprendizaje automático para extraer conocimientos y tomar decisiones informadas. Finalmente, hizo la tarea de supervisar que todos probaran los distintos algoritmos con los diferentes planteamientos de datos y así poder validar cual es el mejor modelo generado.

Juan Manuel Jauregui Rozo (33 puntos)

Rol: Líder del proyecto

Tareas: Se encargó de revisar el documento final entregado, el video entregado, el notebook entregado y la completitud de cada una de las secciones solicitadas para la entrega. Adicionalmente, se encargó de gestionar las reuniones que se realizaron para el desarrollo del proyecto. En cada reunión, se plantearon tareas e hitos que se debían cumplir antes de la siguiente reunión. Este integrante también trabajó el algoritmo de regresión logística y se encargó de redactar y acotar los resultados

finales en el documento y en el video. Este proceso tomó aproximadamente 6 horas distribuidas en los días de desarrollo del proyecto.

Ana Sofía Castellanos Mosquera (33 puntos)

Roles: Líder de Negocio, Líder de datos

Tareas: Se encargó de aplicar en primer lugar, el entendimiento de los datos, para poder aplicar técnicas de limpieza y preprocesamiento. En segundo lugar, implementó diversas técnicas de preprocesamiento, tales como stemming y lemmatizer para determinar que camino de procesamiento llevaba a mejores resultados. Estos datos se dejaron disponibles para la realización del modelo analítico en los archivos EntendimientoDatos.ipynb, que fueron utilizados por cada integrante para la elaboración de los modelos. Por otro lado, en la parte de negocio se encargó de la identificación y descripción de oportunidad de negocio, realizó contacto con el equipo de estadística y cuadró una reunión con el fin de revisar los resultados de esta etapa y verificar el modelo desde pruebas de hipótesis en torno a la exactitud. En lo que respecta al modelo analítico, la estudiante implementó SVM. El tiempo de este proceso fue de aproximadamente 15h, distribuidas a lo largo de la semana santa y semana 9.

7. Entregables

El enlace al repositorio donde se encuentran los entregables es este: https://github.com/acastem15/Inteligencia-de-Negocios/wiki

Para ver en más detalle los tableros de control tanto de entendimiento como de resultados, se puede consultar en GitHub con el nombre: finalTablerosControl.pdf y finalTablerosControl.pbix

8. Referencias

[1] F. Murzone, “Procesamiento de Lenguaje Natural: Stemming y Lemmas,” EscuelaDeInteligenciaArtificial. Accessed: Apr. 06, 2024. [Online]. Available: https://medium.com/escueladeinteligenciaartificial/procesamiento-de-lenguaje-natural-stemming-y-lemmas-f5efd90dca8

[2] A. Jha, “Vectorization Techniques in NLP [Guide],” neptune.ai. Accessed: Apr. 06, 2024. [Online]. Available: https://neptune.ai/blog/vectorization-techniques-in-nlp-guide

[3] Gitlab, ISIS3301, Procesamiento de Textos, Accessed: Apr. 06, 2024. [Online]. Available: https://gitlab.virtual.uniandes.edu.co/ISIS3301/practicas/blob/master/ProcesamientoTextos/Preparaci%C3%B3n_de_textos_estudiante.ipynb

[4] T., Joachims. Text categorization with Support Vector Machines.https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf