Luis Felipe Torres – 202123815

David Pérez Cárdenas – 202123314

Proyecto 1

Inteligencia de Negocios

1) Entendimiento del negocio y enfoque analítico.

Oportunidad/problema Negocio	Entidades como el ministerio de comercio, COTELCO, industria de turismo de Colombia, entre otros, buscan aumentar el turismo en Colombia, por lo tanto, este grupo de entidades tiene distintos objetivos. En primer lugar, ellos esperan que realicemos un análisis de características relevantes de distintos hoteles, las cuales determinan su atractivo en un valor numérico. Esto con el propósito de determinar qué elementos deberían reforzarse para brindar un mejor servicio. Por otro lado, buscan que definamos un modelo que logre determinar la calificación de un turista, dado su reseña, esto con el fin de armar un plan estratégico en pro de aumentar el turismo.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	El proyecto que vamos a realizar lo abordaremos como un problema de análisis de textos con técnicas de aprendizaje automático, ya que el objetivo es poder analizar reseñas para encontrar características positivas y determinar el afecto de una reseña. Alguna de las técnicas que proponemos utilizar son:

	- La vectorización nos permite convertir el texto en vectores numéricos para poder procesarlos más fácilmente. - La técnica de “polynomial features” consiste en elevar características exponencialmente con el fin de mejorar el rendimiento de algunos algoritmos. - La tokenización nos permite dividir una secuencia de texto en tokens (palabras, signos, etc.) para preparar los datos para poder realizar distintas tareas.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Con el desarrollo de este proyecto aquellos beneficiados serían: El ministerio de comercio, industria y turismo, la asociación hotelera y turísitica, cadenas hoteleras de talla Hilton, Estelar, Holiday Inn y hoteles pequeños. Estos actores se verían beneficiados porque, en el caso de un proyecto exitoso, estas entidades podrían generar planes de acción para mejorar sus negocios, dado que, si cuentan con aquellas características importantes para los clientes, y también cuentan con un sistema que determina la calificación dependiendo de una reseña, estos tendrían los insumos suficientes para mejorar ciertos aspectos y realizar estudios más profundos.
Contacto con experto externo al proyecto y detalles de la planeación	Nuestros expertos son: Juan Diego Sevilla, Sara Paiba Vargas y Natalia Cavero. Nos comunicaremos con ellos mediante el correo durante las múltiples etapas de desarrollo, con el fin de que puedan validar el enfoque que le estamos dando al proyecto, el impacto y la presentación de los resultados.

	Finalmente tendremos una reunión el sábado 6 de abril para revisar resultados.

2) Entendimiento y preparación de los datos

Luego de realizar el trabajo anterior, tenemos una mayor comprensión sobre el negocio, las necesidades de este estudio y su propósito. Ahora bien, tenemos que realizar un pequeño análisis sobre los datos para poder comprender que está ocurriendo y que es lo que tenemos que realizar para posteriormente poder realizar unos buenos modelos.

El primer paso por realizar fue ver cómo está organizada la información dentro del "csv" que nos entregaron. Esta información es bastante "simple" teninedo en cuenta que lo único que hay son 2 columnas, donde una de ellas tiene la reseña escrita por el usuario, y la otra las estrellas que este le dio al establecimiento.

Luego decidimos ver la distribución de estrellas para poder determinar si los datos en si nos iban a ser útiles en nuestro análisis, ya que, si el caso era donde una clasificación en específico abarcaba la gran mayoría de datos, esto podía sesgar nuestro modelo ya que no tendría suficiente información como para predecir cuándo una reseña se sale de dicha categoría.

Otras cosas que decidimos mirar fueron, si había existencia de datos vacíos y la distribución de palabras en una reseña. Estos elementos nos ayudaron a comprender un poco mejor los datos.

Para la preparación de los datos lo primero que decidimos realizar fue una limpieza general teniendo en cuenta algunos aspectos irrelevantes como las letras tildadas y las mayúsculas.

Posteriormente realizamos un proceso de vectorización de los datos mediante el uso de un pipeline con ayuda de la función "TfidfVectorizer". Esta función nos ayuda en múltiples aspectos, en primera instancia si los datos no están tokenizados realiza este paso, además decide contar cuantas veces aparece cada toquen y le asigna una frecuencia. Posteriormente realiza un cálculo de la frecuencia inversa para poder medir y comparar ambas métricas. Teniendo en cuenta esto, el logra devolver una matriz donde cada columna es la palabra y las métricas encontradas para cada una.

3) Modelado y evaluación

Luego de haber realizado la etapa anterior, empieza ahora la creación de modelos y la evaluación de cada uno de ellos.

Los modelos que escogimos fueron los siguientes: RidgeClassifier, LogisticRegression y ComplementNB.

“Ridge Classifier”, este clasificador utiliza la regresión de Ridge. Este método Ridge permite estimar los coeficientes de modelos de regresión en múltiples escenarios donde las variables independientes están altamente correlacionadas. Aplica regularización L2 para reducir el sobreajuste en el modelo y manejar la multicolinealidad (alta correlación entre características).

“Logistic Regression”, este es un modelo estadístico que modela las variables con coeficientes para cada característica (estos coeficientes se estiman utilizado la información de entreno), luego realiza unos cálculos de la suma ponderada de los coeficientes encontrados, para luego pasarlos por una función lógica y producir valores de 0 a 1 (que determina la probabilidad de la instancia estar en la categoría). Posteriormente, dependiendo del número resultado de cada categoría se determina si el acumulado de elementos cumplen con el valor predicho.

“ComplementNB”, este modelo es una adaptación del algoritmo MNB (multinomial naive Bayes) que es usado para data sets no balanceados. Este modelo logra determinar valores más adecuados cuando los datos estan más sesgados hacia una dirección.

Fue fundamental escoger estos 3 modelos ya que cada uno lograba abarcar un ámbito distinto. “RidgeClassifier” nos permitía ver como se comportaban estos datos cuando aplicábamos modelos de tipo regresión, y por su parte, este modelo logra utilizar múltiples modelos de regresión donde la correlación entre algunas características y el resultado era alta, por otro lado el modelo “LogisticRegression” nos permite hacer un enfoque más estadístico a la importancia de las categorías y finalmente “ComplementNB” nos permitía abarcar el escenario de un sesgo en los datos. En conclusión, utilizar estos tres diferentes modelos fue bastante adecuando, ya que cada uno de ellos tenía un acercamiento diferente a los otros, y nos permitió abarcar más escenarios.

4) Resultados

Algoritmo	Exactitud	Precisión	Recuperación	F1
Ridge Classifier	0.462	0.451	0.462	0.451
Logistic Regression	0.484	0.476	0.484	0.475
Complement NB	0.487	0.468	0.487	0.468

Podemos ver que los 3 algoritmos tienen valores similares para sus diferentes métricas:

- Exactitud: Esta métrica indica que porcentaje de las predicciones que realizo el modelo corresponden a la clase correcta.
- Precisión: Esta medida indica que porcentaje de las predicciones son realmente correctas.
- Recuperación: Indica el porcentaje de los casos positivos reales que el modelo detecta.
- F1-Score: Es una forma de combinar precisión y recuperación en una sola medida, por lo que nos resulta útil cuando necesitamos un equilibrio entre estas 2 métricas.

Al analizar los resultados de los algoritmos Ridge Classifier, Logistic Regression y Complement NB, se observa que presentan valores similares en términos de exactitud, precisión, recuperación y F1-Score. Esto sugiere que no hay diferencias significativas en el desempeño entre estos tres algoritmos para este problema en específico.

Sin embargo, encontramos que el modelo que mejor se acerca a lo que buscamos es el Logistic Regression.

(El video se encuentra en la wiki)

5) Mapa de actores relacionado con el producto de datos creado

6) Trabajo en equipo

Este trabajo realizado fue labor de Luis F. Torres y David Pérez. Ambos integrantes trabajaron (cada uno) alrededor de 8-10 horas en el proyecto, aunque hayan tenido roles diferentes, hubo muchas instancias donde el trabajo en equipo era fundamental, por esta razón cada integrante debería de tener 50 puntos (que suman 100 teniendo en cuenta que fueron 2 integrantes).

El estudiante Luis F. Torres tenía el rol de líder de modelado y líder de análisis, este estudiante se encargaba de realizar las labores que implicaban realizar acciones con los datos y posteriormente utilizarlos en diferentes modelos. Posteriormente este estudiante se encargó de realizar el dicho análisis necesario para poder escoger aquel modelo que era más pertinente para conseguir los objetivos del negocio. Se destaca su esfuerzo y su ingenio para poder escoger el mejor paso a seguir para lograr sus propios objetivos.

El estudiante David Pérez tenía el rol de líder de planeación y líder de negocio, este estudiante se encargaba de realizar labores que implicaban planear el paso a seguir del grupo, determinar que opción tenía que ser tomadas en ciertos escenarios durante el proceso de modelado. De forma inicial este estudiante se encargó de

realizar un análisis profundo sobre el negocio, las necesidades de este y el enfoque que debería tomar el proyecto. Se destaca su esmero y cuidado a la hora de definir acciones como grupo.

De forma general, el proyecto realizó alrededor de 3 reuniones principales. La primera fue necesaria para realizar un entendimiento del caso de estudio y definir un plan de acción general para el proyecto. La siguiente reunión tuvo el enfoque de realizar el desarrollo del proyecto como tal, escogiendo que modelos serían utilizados y el análisis de lo que estos determinan. Finalmente, ocurrió la última reunión donde se realizó un análisis de resultados, se verificó el trabajo y se llegó a un consenso de que elementos eran necesarios para mejorar.

Aquellos elementos que se resaltaron para mejorar son: Realizar una comunicación más efectiva; esto no implica que haya habido mala comunicación, sino que debería priorizarse su efectividad, y también se encontró que otro aspecto a mejorar es la definición de prioridades; no fue del todo correcta la decisión de prioridades en el proyecto, eso es un aspecto para mejorar.

7) Sustentación y evaluación del aporte individual

Esto se encuentra en la wiki