

**UNIVERSIDAD DE LOS ANDES  
DEPARTAMENTO DE INGENIERIA DE  
SISTEMAS Y COMPUTACIÓN**



**PROYECTO 1: ETAPA 2**

**ISIS 3301 – INTELIGENCIA DE NEGOCIOS**

**JUAN CAMILO ORTIZ**

**Grupo 27:**

Tomas Ángel – 202020366

Raúl Rincón – 202120414

Luis Felipe Dussán – 201912308

## Tabla de Contenidos

Tareas asignadas:.....	3
<i>Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API.....</i>	<i>3</i>
Desarrollo de la aplicación y justificación.....	4
Herramientas Utilizadas .....	7
<i>Resultados .....</i>	<i>7</i>
<i>Trabajo en equipo.....</i>	<i>10</i>

## Tabla de Figuras

Ilustración 1 Fragmento del código donde se exporta el modelo.....	4
Ilustración 2 Funcionamiento app web .....	5
Ilustración 3 Funcionamiento API Rest .....	5
Ilustración 4 Muestra de los resultados utilizando la aplicación web.....	9
Ilustración 5 Muestra de la información de las reseñas usando la aplicación web .....	9
Tabla 1 Actores finales.....	7

Tareas asignadas:

Luis Felipe Dussán

- Ingeniero de datos
- Redacción del documento

Raul Santiago Rincon

- Ingeniero de datos
- Ingeniero de software responsable de diseño de la aplicación y resultados
- Ingeniero de software responsable de desarrollar la aplicación final
- Redacción del documento

Tomas Ángel

- Líder de proyecto
- Ingeniero de datos
- Ingeniero de software responsable de diseño de la aplicación y resultados
- Redacción del documento

Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API

Para esta etapa del proyecto se describe el proceso de implementación de un sistema automatizado para el análisis de sentimientos en reseñas turísticas. Con base en la anterior etapa, se determinó que el mejor modelo de machine learning para análisis de textos entre “*Logistic Regression*”, Naive Bayes, y KNN (K nearest neighbor) fue el de Naive Bayes con el más alto F1 score (0.47). Determinamos esta métrica para escoger el modelo puesto que es efectiva y equilibrada teniendo en cuenta la precisión y el recall.

Para que el estudio realizado anteriormente sea de especial utilidad para los diferentes interesados en el proyecto, determinamos que realizar un pipeline era la mejor forma de automatizar este proceso para predecir la calificación de una reseña. Se busca desplegar un ambiente de producción para acceder con un API y que los usuarios puedan tener una forma fácil y amena de ver estos datos.

El primer paso fue limpiar y preparar los datos de las reseñas para tokenizar y lematizar las palabras. Al tokenizar se divide un texto completo en palabras separadas para que de esta forma sea más fácil eliminar palabras y caracteres que no son valiosas en el modelo. Por otro lado, lematizar es un proceso lingüístico que consiste en reducir las palabras a su forma base permitiendo tratar diferentes formas de una palabra como si fuera la misma. Realizar este proceso es importante para normalizar todas las reseñas y luego pasarlo por el modelo de Naive Bayes que fue el segundo paso a realizar para la automatización de todo. El tercer paso después de entrenar el

modelo fue guardar el modelo realizado con los datos de prueba (o datos no etiquetados), para un posterior uso en producción. En nuestro caso, guardamos el modelo dentro de un archivo *.joblib* puesto que esto nos dará facilidad en el uso de la herramienta *fastAPI*, que es la herramienta donde se desplegara la aplicación, siendo este último el cuarto y último paso.

Decidimos utilizar *fastAPI* porque es un framework web de Python que ofrece beneficios significativos para el desarrollo de aplicaciones API y es altamente reconocido por su rendimiento, comparable incluso con otros frameworks como Node.js y Go. Su capacidad para integrar con Pydantic mejora la mantenibilidad del código y reduce errores, factores cruciales para personas que no sean muy hábiles a nivel tecnológico. Asimismo, implementamos Bootstrap como framework con el propósito de garantizar una experiencia de usuario consistente, atractiva y adaptable.

En el primer paso del proceso, nos enfocamos en comprender los datos: cómo están organizados, la distribución de las calificaciones en las diferentes reseñas, la presencia de datos duplicados o nulos, y el formato en el que se guarda la información. Al analizar estos aspectos, se observó que las calificaciones siguen un patrón lineal, donde las calificaciones de 1 son las menos frecuentes en los datos y las de 5 son las más comunes. Esto será importante para la parte de resultados y el entrenamiento del modelo que se realice.

La segunda parte del proceso consistió en preparar los datos para el modelo, así que se decidió tener todas las palabras en una reseña de la forma más simple para realizar la vectorización de estos. Esto implicó eliminar caracteres que no fuera alfanuméricos, y para las palabras que tuvieran tildes y otros caracteres especiales se les removían dichos caracteres para que las palabras quedaran en minúscula.

Además de la tokenización y lematización, fue necesario realizar una vectorización sobre las reseñas puesto que pasa las palabras a números para que los algoritmos funcionen. En la primera etapa del proyecto se decidió que se utilizaría el algoritmo de Naive Bayes como modelo de análisis de textos con la vectorización de count vectorizer.

Lo anterior permitió ir al cuarto paso que fue la creación del pipeline, donde se automatizaron todos los anteriores pasos para que el modelo se pueda utilizar en cualquier momento. Para facilitar el uso de este modelo previamente entrenado de clasificación de reseñas, se realizó una aplicación web donde los diferentes usuarios puedan cargar archivos de reseñas y se muestra la calificación para cada reseña.


```
# Exportar el pipeline a un archivo .joblib
dump(pipeline, '/Users/tomasangel/Documents/GitHub/proyecto1etapa2/data/modelo.joblib')
```

*Ilustración 1 Fragmento del código donde se exporta el modelo*

## Desarrollo de la aplicación y justificación

Bienvenido al proyecto

Turismo los Alpes



Acá podrás encontrar una plataforma para subir las reseñas y poder hacer una predicción de la calificación de cada una de las reseñas.

Debes subir un documento en formato CSV con las reseñas y posteriormente dar click a subir

Seleccionar archivo

Ninguno archivo selec.

Subir CSV

Para visualizar las reseñas puedes dar click acá abajo

Leer reseñas

Después de subir tus datos puedes hacer predicción de la calificación que tendrías las reseñas adjuntadas

Hacer predicción

Visualizar resultados de la predicción

Ver Resultados

*Ilustración 2 Funcionamiento app web*

```

INFO: 127.0.0.1:63741 - "GET /static/barplot.png HTTP/1.1" 304 Not Modified
INFO: 127.0.0.1:63741 - "GET /static/scatterplot.png HTTP/1.1" 304 Not Modified
INFO: 127.0.0.1:63741 - "GET / HTTP/1.1" 200 OK
INFO: 127.0.0.1:63741 - "GET /tablero HTTP/1.1" 200 OK
INFO: 127.0.0.1:63741 - "GET /static/barplot.png HTTP/1.1" 304 Not Modified
INFO: 127.0.0.1:63741 - "GET /static/scatterplot.png HTTP/1.1" 304 Not Modified
INFO: 127.0.0.1:63741 - "GET / HTTP/1.1" 200 OK

```

*Ilustración 3 Funcionamiento API Rest*

Antes de desarrollar la aplicación, se debió considerar la participación de diferentes actores y usuarios, ya que serían quienes la utilizan. Para esto, nos apoyamos del grupo de estadística con el que trabajamos para realizar la revisión de la tabla de beneficiarios de todo el proyecto. Para la primera parte del proyecto se realizó una tabla preliminar con los que considerábamos los actores más importantes en el proyecto y sus respectivos roles, pero después de una revisión más detallada por el grupo de estadística, nos recomendaron fuertemente realizar cambios para los roles de turista y el Ministerio de Comercio, Industria y Turismo en Colombia. Para el primer rol, consideraban que, además de ser un actor Usuario-Cliente, también serían beneficiarios al revisar las reseñas anteriores para saber que sitios eran buenos y malos al realizar una visita turística. Para el segundo rol, argumentaban lo mismo, poniéndolos como beneficiarios dado que, al mejorar los servicios turísticos del país, se aumentan las visitas turísticas y mejora la reputación de este. Teniendo lo anterior en consideración, el mapa de actores final relacionada con el producto de datos creados se vería del siguiente modo:

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
<b>Ministerio de Comercio, Industria y Turismo en Colombia</b>	Financiador	Puede utilizar la información para desarrollar políticas y estrategias que promuevan el	Si el modelo no tiene un buen desempeño, las políticas y estrategias pueden basarse

		turismo en Colombia.	en información incorrecta, lo que podría no tener el efecto deseado en el turismo. Adicionalmente, si no se realiza un buen modelo, se podrían realizar malas inversiones.
<b>Ministerio de Comercio, Industria y Turismo en Colombia</b>	Beneficiario	Puede utilizar la información para mejorar los servicios turísticos del país.	Si el modelo no tiene un buen desempeño y se realizan cambios innecesarios en los servicios turísticos, la reputación se podría ver afectada y por ende no atraería turistas al país.
<b>Asociación hotelera y turística de Colombia (COTELCO)</b>	Usuario-Cliente	Puede utilizar los resultados para mejorar la calidad de los servicios turísticos y promover destinos turísticos específicos.	Si el modelo no funciona correctamente, las estrategias de promoción podrían dirigirse incorrectamente, lo que podría afectar la reputación de los destinos turísticos.
<b>Cadenas hoteleras (Hilton, Hoteles Estelar, Holiday Inn, etc.)</b>	Beneficiario	Pueden utilizar la información para adaptar sus servicios y promociones a las preferencias de los turistas	Si el modelo no es preciso, las adaptaciones de servicios y promociones podrían no ser efectivas, lo que podría afectar su competitividad en el mercado.
<b>Hoteles pequeños en diferentes municipios de Colombia</b>	Beneficiario	Pueden mejorar sus servicios y promocionarse de manera más efectiva, lo que puede aumentar su popularidad entre los turistas.	Si el modelo no funciona correctamente, las mejoras en los servicios y la promoción podrían no ser adecuadas, lo que podría no

			tener el efecto deseado en la atracción de turistas.
<b>Turistas locales y extranjeros</b>	Usuario-Cliente	Pueden beneficiarse al recibir recomendaciones más precisas y adaptadas a sus preferencias individuales.	Si el modelo no es preciso, las recomendaciones pueden no ser relevantes para los turistas, lo que podría afectar su experiencia de viaje.

*Tabla 1 Actores finales*

## Herramientas Utilizadas

### **FastApi:**

El uso de FastAPI lo podemos argumentar por estos puntos:

**1. Rendimiento y eficiencia:** En el contexto de nuestra aplicación basada en aprendizaje automático que requiere un alto rendimiento, FastAPI puede manejar eficientemente una gran cantidad de solicitudes concurrentes, lo que la hace adecuada para escenarios donde se necesite procesar datos en tiempo real o en lotes.

**2. Facilidad de implementación de API:** FastAPI facilita la creación de una API web al proporcionar una sintaxis simple y clara para definir endpoints y modelos de datos. Esto permite a los desarrolladores implementar rápidamente una interfaz atractiva y fácil de usar para interactuar con el modelo analítico, sin tener que preocuparse demasiado por la infraestructura subyacente.

**Joblib:** La usamos para el almacenamiento y la carga eficientes de objetos Python en archivos.

**Bootstrap:** Implementamos Bootstrap como framework con el propósito de garantizar una experiencia de usuario consistente, atractiva y adaptable.

## Resultados

En este proyecto, desarrollamos una aplicación web utilizando el marco de trabajo *FastAPI* en Python, con el objetivo de permitir a los usuarios interactuar con un modelo de aprendizaje automático de manera eficiente y efectiva. Los principales resultados obtenidos son los siguientes:

- **Desarrollo de la Aplicación Web:** Utilizando *FastAPI*, creamos una API web que proporciona endpoints para la interacción con el modelo de aprendizaje automático. Esto incluye la definición de endpoints para enviar datos al modelo,

recibir predicciones y resultados, y gestionar archivos cargados por los usuarios.

- **Integración del Modelo de Aprendizaje: Automático:** Implementamos la lógica necesaria para integrar el modelo de aprendizaje automático en la aplicación web. Esto incluye la carga y preprocesamiento de datos, la ejecución de predicciones utilizando el modelo entrenado y la generación de respuestas adecuadas para los usuarios.
- **Interfaz de Usuario Interactiva:** Desarrollamos una interfaz de usuario interactiva utilizando archivos estáticos, como HTML, CSS y JavaScript, servidos por la aplicación *FastAPI*. Esto permite a los usuarios interactuar fácilmente con la aplicación a través de un navegador web, cargando archivos, enviando datos y recibiendo resultados.
- **Página de Inicio:** La página de inicio ofrece una introducción concisa a nuestro proyecto "Turismo los Alpes", proporcionando información sobre las funciones disponibles en la aplicación. Los usuarios pueden acceder fácilmente a las acciones principales de nuestra aplicación, como *Leer reseñas* y *Subir CSV*, a través de enlaces y formularios intuitivos.
- **Visualización de Reseñas:** Implementamos una página dedicada a la visualización de reseñas ingresadas por los usuarios. Utilizando un formato de tabla, se presenta la información sobre las reseñas existentes. Esta funcionalidad permite a los usuarios revisar y analizar las reseñas de manera eficiente.
- **Funcionalidad de Subir Archivo CSV:** Desarrollamos un formulario que permite a los usuarios cargar archivos CSV conteniendo reseñas. Esta funcionalidad facilita la entrada masiva de datos, optimizando el proceso de ingreso de información relevante para el proyecto. Una vez cargados, los datos se procesan automáticamente para luego mostrarse en la página de visualización de reseñas.
- **Predicción:** Mediante un botón presentamos la predicción que realizó el modelo, esto se realiza de forma intuitiva y fácil para el usuario.
- **Processing:** Realizamos un procesamiento sobre las predicciones obtenidas, lo cual nos permite destacar las palabras más relevantes y significativas dentro de las reseñas. Esta herramienta brinda a nuestros clientes la capacidad de identificar rápidamente los aspectos más relevantes y críticos expresados por los usuarios en sus comentarios. Al mostrar las palabras clave y su frecuencia de aparición, facilitamos la toma de decisiones de negocio al proporcionar insights claros y concisos sobre las opiniones y percepciones de los usuarios. Estos análisis permiten a nuestros clientes comprender mejor la satisfacción del cliente, detectar áreas de mejora y tomar medidas estratégicas para optimizar sus productos o servicios, así como mejorar la experiencia general del cliente.

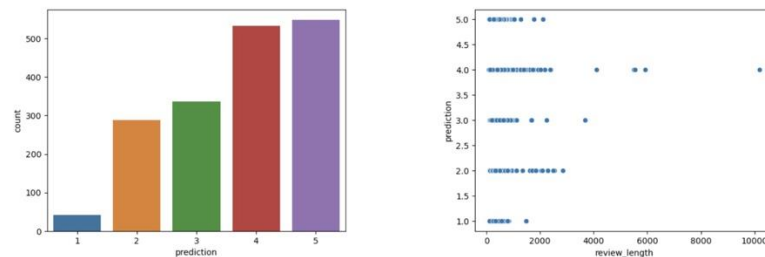


## Resultados de la Predicción del Modelo

Texto	Calificación
La primera noche nos encontramos en la habitación con un nido (5) de cucarachas muertas y la alfombra muy sucia...que por cierto nunca limpiaron hasta que reclamamos. La cena de fin de año fue un bufete que acabó en un tremendo desorden por el excesivo número de mesas vendidas, no se podía caminar para ir a servirte la comida, para luego encontrarte con bandejas vacías, mi esposa y yo y otra pareja "NO CENAMOS" porque nos encontramos con puras bandejas vacías!!...pagamos \$2800 pesos por por pareja por nada!! El servicio del restaurante pésimo, pésimo,pésimo...se tardaron muchísimo tiempo en atendernos, nos trajeron el desayuno en partes, nunca nos dieron el café que pedimos desde que llegamos...un pésimo servicio El frigobar vacío y después de reclamar lo surtieron pero_x0085_	2
A una calle de catedral con platillos tradicionales, tipo Gourmet, de buen sabor y calidad. Si bien ofrecen una carta con los platillos más representativos de cada temporada como mole de caderas, poblano, chinicuiles, chiles en nogada, escamoles y cemitas; son principalmente para degustación, no...Más	5
Porciones muy miserables Agua con sabor a cloro Muy distraídos los meseros No llena las expectativas Hay mejores opciones	2
Cartagena encanta. De todas las ciudades coloniales que hemos visitado es la más linda y mejor conservada. Recomiendo recorrerla tanto de día como de noche ya que son visiones distintas que vale la pena vivir. Llena de vida, tiendas, restaurantes, colorida, vibrante, calles y rincones que en cada momento guardan una sorpresa. Mi calificación habría sido excelente sino fuera por algunos detalles que deberían mejorar. Lamentablemente falta preocupación por la limpieza de las calles (problema que se repite en el resto de la ciudad) y un mejoramiento en los alcantarillados de las aguas servidas. Además por la estrechez de las calles deberían habilitar vías exclusivas para peatones.	5
Ibamos con mucha ilusión de disfrutar el espectáculo de luz y sonido pero la verdad nos decepcionó,te ubican en una esquina a la entrada de las ruinas,y no al centro por lo que no puedes apreciar casi nada,las personas se paran a tomar fotografías debido a que no se alcanza a ver nada si estas sentado,al inicio y al final del espectáculo no dejan prendidas las luces para que puedas tomar fotografías, al	3

Ilustración 4 Muestra de los resultados utilizando la aplicación web

- **Graficas:** Mediante otro botón el usuario puede navegar hacia graficas correspondiente a nuestro modelo analítico, de esta forma el cliente puede sacar insights de esta e información adicional.



## Palabras importantes

Calificación	Palabras mas comunes en las reseñas
1	hotel, habitacion, si, mas, habia, solo, agua, mal, servicio, nunca, bien, personal, peor, three, recepcion, hacer, minutos, nadie, luego, dia
2	hotel, habitacion, mas, comida, servicio, si, habia, personal, habitaciones, solo, mal, noche, bien, dos, dia, lugar, recepcion, agua, restaurante, desayuno
3	hotel, mas, si, lugar, servicio, bien, habitacion, comida, solo, buena, personal, habitaciones, mejor, bastante, habana, ser, ver, noche, bueno, restaurante

Ilustración 5 Muestra de la información de las reseñas usando la aplicación web

- **Navegación y Experiencia del Usuario:** Incorporamos una navegación intuitiva que permite a los usuarios moverse fácilmente entre las diferentes secciones de la aplicación. Esto contribuye a mejorar la experiencia del usuario, asegurando una navegación fluida y sin inconvenientes.

Como último paso del proceso y creación del proyecto, le mostramos la aplicación final a los estudiantes de estadística con los que trabajamos para este proyecto y sus comentarios fueron los siguientes:

- Es una aplicación muy útil para entender las necesidades de un negocio para ver sus fortalezas y debilidades, especialmente en la industria hotelera.
- Es bueno que el programa de las predicciones y se puedan ver tablas y graficas relacionadas con esta información que me ayuden a entender de mejor manera el archivo de reseñas que se generó.
- Hubiera sido bueno

## Trabajo en equipo

### Reuniones con el grupo de estadística

Creamos el espacio para 4 reuniones, estas se llevaron a cabo:

- Viernes 12 de abril: Primera presentación del proyecto
- Martes 16 abril: Segunda revisión de la presentación y el proyecto teniendo en cuenta las recomendaciones dadas en la primera parte.
- Viernes 19 de abril: Penúltima revisión del proyecto. Se detallan los últimos detalles que debería tener el proyecto.
- Sábado 20 abril: Ultima revisión del proyecto y consolidación del trabajo. Se muestra la aplicación y se recibe la retroalimentación para terminar el trabajo

Durante nuestras sesiones con los estudiantes de estadística, recibimos valiosas opiniones y comentarios sobre diversas áreas de nuestro proyecto. Estos incluyeron la calidad de nuestras presentaciones, la estructura del modelado de datos, definición de actores, la claridad en la presentación de los resultados y observaciones adicionales sobre el modelo empleado. La participación de los estudiantes de estadística en conjunto fue efectiva y útil, ya que nos proporcionó orientación y dirección para mejorar y fortalecer nuestro proyecto.

**Número de horas dedicadas:** 3 horas por semana para la construcción de la integración del modelo y construcción de la interfaz.

**Retos enfrentados y soluciones:** Integración del modelo con la interfaz, fue retador realizar la construcción en *FastAPI*, y alimentar el modelo con nuevas reseñas.

**Puntos de mejora:** Ser más proactivos en la recopilación de Feedback de usuarios o colegas sobre la interfaz y la funcionalidad del modelo antes de proceder con su construcción. Aunque recibimos comentarios de los estudiantes de estadística, reconocemos que debimos haber presentado un mockup de la interfaz al cliente antes de iniciar la construcción. Construir primero y validar después con los estudiantes de estadística resultó en un proceso menos eficiente, ya que se identificaron numerosas correcciones que podrían haberse evitado con una revisión preliminar por parte del cliente.