Andrés Takashi Bravo Yaguchi – 201911714

Sara Plazas Copete – 201821723

Kevin Steven Gámez Abril – 201912514

**PROYECTO I – ETAPA II**

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

Para automatizar la preparación de datos, construcción del modelo, persistencia de este y acceso por a este por medio de un API hemos creado un pipeline de procesamiento de lenguaje natural utilizando Python. En primer lugar, se carga en un   
DataFrame de la librería pandas el archivo csv donde se encuentran las reseñas de las películas clasificadas por un experto. A este DataFrame se le eliminan las reseñas que se encuentren repetidas y se eliminan las reseñas que se encuentran en un idioma diferente al español debido a que al negocio solo le interesan las reseñas en este idioma.

Como se explicó en la etapa anterior del proyecto, se realizó la preparación de datos para entrenar un modelo de análisis de sentimientos a partir de reseñas de películas en español. En este caso para la elaboración de un modelo y la persistencia de este decidimos usar pipelines de la librería de Python Scikit Learn, en donde definimos una clase llamada TextPreprocessor donde se realizan diversas transformaciones a las reseñas, incluyendo la eliminación de entradas duplicadas y reseñas en otros idiomas, así como la tokenización, lematización y eliminación de palabras que no aportan información valiosa para el análisis. También se hizo uso de la librería num2words para transformar los números en su representación en texto. Finalmente, se utilizó el modelo Tf-Idf para convertir las reseñas en una matriz numérica que considera la frecuencia relativa de cada palabra en las reseñas. Con estas transformaciones, los datos se encuentran listos para entrenar un modelo de clasificación de textos.

El modelo de clasificación escogido fue Regresión logística debido a que fue el que obtuvo mejores métricas en las pruebas realizadas. El pipeline se define mediante la función make\_pipeline de la librería scikit-learn. Este pipeline tiene dos etapas: la primera etapa es un ColumnTransformer que utiliza la clase TextPreprocessor para preprocesar los datos de texto y convertirlos en una representación vectorial mediante la técnica TF-IDF. La segunda etapa es un modelo de regresión logística que se entrena utilizando los datos preprocesados. Finalmente, se utiliza la librería Dill para guardar el modelo entrenado en un archivo joblib para su posterior uso en la construcción del API. Dill es una alternativa a pickle y al método dump de joblib que permite serializar y des serializar objetos Python más complejos de los que soporta pickle y dump.

# Desarrollo de la aplicación

Una vez construido el API con el pipeline del modelo de clasificación seleccionado en la etapa anterior, pasamos a la creación de la aplicación que permitirá la interacción con el usuario. Se creó una aplicación React para mayor facilidad y con ayuda de diferentes librerías, como Bootstrap, se fueron implementando todos los componentes necesarios para que un usuario dado pudiese interactuar con el modelo, ingresando nuevas reseñas y recibiendo como respuesta su clasificación. Se crearon dos funcionalidades importantes para esto, y es que se quería habilitar la aplicación para recibir tanto reseñas individuales como conjuntos más grandes de reseñas, lo anterior con el fin de asegurar un servicio cómodo y completo para el usuario. Asi pues, en la aplicación se integra tanto un cuadro de texto para que un usuario pueda ingresar solo una reseña a la vez, como un botón que posibilita cargar un archivo de tipo csv con varias reseñas. Estas entradas luego serian enviadas al backend para realizar la predicción con ayuda del pipeline y devolver la predicción generada a la aplicación. Sin embargo, cabe resaltar que es de gran importancia el cómo se presenta esta información al usuario de forma amigable y organizada, en especial en el caso de ingresar un conjunto de varias reseñas. En este punto la colaboración con el experto de estadística fue de gran ayuda pues este nos asesoró sobre la información estadística más relevante para presentar y como mostrarla. Después de esta asesoría se optó por apoyarse en herramientas como tablas, filtros y gráficas de pastel, entre otras. Lo anterior fue de gran utilidad para asegurar mayor comodidad y facilidad para el usuario.

# Definición de usuarios y justificación

Ahora bien, para la justificación de la aplicación es de gran importancia pensar “¿Quién la podría usar y para que le podría servir?”. La aplicación de análisis de sentimientos de películas puede ser beneficiosa para diferentes tipos de usuarios, como se mencionó en la anterior etapa del proyecto, el uso se podría extender a varios contextos en los que usuarios y trabajadores de la industria del entretenimiento pueden hacer uso de esta herramienta. Con ayuda del experto, concluimos que algunos potenciales usuarios son:

* Estudios de cine: La aplicación les ayudaría a obtener información sobre la calidad y popularidad de sus películas, con esto pueden lograr un ajuste en sus estrategias de marketing y producción.
* Investigadores de mercado: En este caso, los investigadores de mercado haciendo uso de la aplicación pueden llevar a cabo análisis de del mercado y obtener datos de gran valor sobre las preferencias de la audiencia, ya sea a nivel internacional, o local con grupos demográficos específicos.
* Críticos de cine y periodistas: Haciendo uso de la aplicación podrán obtener información sobre la percepción de los consumidores sobre una película y proporcionar una crítica más informada y precisa.
* Audiencia: En cuanto a los usuarios de las plataformas de entretenimiento, la aplicación tiene el potencial de mejorar la precisión y eficacia de la toma de decisiones al permitir que se tomen en cuenta las opiniones de otros espectadores al momento de elegir qué películas ver o evitar. En consecuencia, se puede incrementar el nivel de fundamento en la toma de decisiones en este ámbito mediante el uso del API.

Pasando a un análisis más específico y detallado, se determinó una empresa de una industria especifica en la que esta se podría utilizar, y a partir de esto definimos dos roles que se verían beneficiados por un modelo de análisis de sentimientos con el fin de adecuar la aplicación de la manera más apropiada posible. Específicamente, decidimos centrar el desarrollo pensando en el productor ejecutivo y la gerente de mercadeo de una empresa productora de películas para ayudarlos en los procesos de distribución y mercadeo de películas. El director ejecutivo de una productora controla, y a menudo proporciona, el financiamiento de las películas que se desarrollan por la empresa y, por lo tanto, es importante para esta persona conocer la recepción que están teniendo las películas para decidir cuales proyectos financiar, cuales rechazar y como distribuir correctamente los recursos de publicidad entre los proyectos aceptados. Teniendo en cuenta esto, un director ejecutivo podría beneficiarse de la aplicación al tener la posibilidad de analizar automáticamente conjuntos grandes de reseñas de películas y determinar si una película tiene mayormente comentarios negativos o positivos por la audiencia; con esta información podría decidir cortar el financiamiento a proyectos que estén teniendo mayormente reseñas negativas o aumentar el financiamiento a proyectos que muestren mejores rendimientos. Por otro lado, la gerente de mercadeo de una productora se encarga de diseñar y ejecutar estrategias para promocionar las películas que son producidas por la empresa. Consecuentemente, podría utilizar la aplicación para identificar el impacto que está teniendo una estrategia especifica en los consumidores y decidir si continuar la campaña publicitaria o cambiar de estrategia según el porcentaje de reseñas positivas que se observen.

A continuación, presentamos la validación del proyecto que realizó el experto en estadística que consultamos. En esta sección, el experto explica los criterios que considera importantes para evaluar el modelo y da su opinión sobre los objetivos alcanzados con el desarrollo.

# Validez de los resultados (sección escrita por el experto en estadística)

En el contexto de los estudios estadísticos, resulta crucial prestar atención a varios factores determinantes, entre los que destacan el tamaño de la muestra, para este contexto la polaridad (sentimiento negativo o positivo), la técnica de muestreo empleada, y, por supuesto, el análisis estadístico utilizado. Estos aspectos son vitales para poder evaluar la fiabilidad y validez de las conclusiones extraídas y, en última instancia, respaldar la toma de decisiones informadas.

**Método de muestreo:** El método de muestreo simple es una técnica estadística ampliamente utilizada y altamente confiable para realzar estudios. Esta técnica se basa en la idea de que todos los individuos de una población tienen la misma probabilidad de ser elegidos para formar parte de la muestra, lo que asegura una muestra representativa y precisa al utilizar este método, se aumenta significativamente la probabilidad de obtener resultados precisos y confiables que pueden ser generalizados con precisión de toda la población.

**Tamaño de la muestra:** En función de los objetivos del estudio y la población en cuestión, la selección de un tamaño de muestra óptimo es de vital importancia en el diseño de investigaciones estadísticas.

En algunos casos, se ha demostrado que un tamaño de muestra significativa proporciona la precisión necesaria en los resultados, aunque el tamaño de la muestra ideal puede variar según el contexto. Por eso, es importante tener en cuenta que el aumento del tamaño de la muestra suele aumentar la precisión, aunque también aumentará los costos y el tiempo necesario para llevar a cabo el estudio. Por lo tanto, es esencial equilibrar la precisión deseada con los costos y el tiempo disponibles.

En cualquier caso, la muestra debe ser representativa de la población objetivo, y el tamaño de muestra debe ser suficiente para garantizar la precisión de los resultados. Con base en estos criterios, se puede justificar el uso de una muestra que en este caso se tomó de 5000 datos para el entrenamiento del modelo. Esto permite entender que los resultados solo serán una muestra y no representan necesariamente la opinión de toda la población. Sin embargo, pueden ofrecer una compresión útil de las opiniones generales del público hacía una película, como sucede en este caso.

**Medida del sentimiento:** En este estudio, se midió la actitud hacía una película mediante la división de dos categorías: positiva y negativa. Como resultado, el análisis estadístico basado en la frecuencia resulta adecuado para evaluar los resultados. Este análisis estadístico permitiría la evaluación del número de veces que se seleccionó cada una de las dos categorías en la muestra.

Además, es de destacar que el análisis de correlación para examinar la relación entre la actitud positiva y negativa proporciona información muy valiosa sobre la percepción de la audiencia con relación a un producto, servicio o película en particular. Es importante tener en cuenta que la elección de la medida se basa en los objetivos del estudio y en la información que se pretende obtener, tal como se ha definido en el proyecto.

**Análisis estadístico:** En el proyecto que se llevó a cabo, se aplicaron dos técnicas de análisis estadístico para evaluar variables categóricas o nominales, incluyendo la medición del sentimiento expresado por el público mediante opciones positivas y negativas.

La técnica de análisis de frecuencia se utilizó para determinar la frecuencia de cada opción seleccionada dentro de la muestra, lo que proporcionó información relevante sobre la opinión del público respecto al objeto de estudio. Asimismo, se empleó un análisis de correlación para evaluar la relación entre el sentimiento positivo y negativo con otras variables de interés, lo que permitió obtener información adicional sobre la percepción del público y su relación con otros factores relevantes para el estudio.

En cuanto a la representación de los datos, se utilizó un diagrama circular para mostrar con porcentajes de la muestra el sentimiento hacia un comentario o muchos comentarios contenidos en un documento.

En conclusión, la aplicación de técnicas de análisis estadístico para evaluar variables categóricas o nominales, como la medición del sentimiento expresado por el público, resultó ser una estrategia adecuada para nuestro proyecto. La combinación de la técnica de análisis de frecuencia y el análisis de correlación, junto con la representación de los datos mediante un diagrama circular, permitió obtener una visión clara y precisa de la percepción del público sobre el objeto de estudio.

Es evidente que el uso de estas técnicas estadísticas brindó información relevante y valiosa para la empresa interesada, lo que demuestra que el proyecto cumplió con éxito con sus objetivos y metas establecidas. En definitiva, la aplicación de técnicas estadísticas adecuadas es fundamental para obtener resultados precisos y fiables que permitan tomar decisiones acertadas en cualquier empresa.

# Trabajo en equipo

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Integrante** | **Rol** | **Horas** | **Tareas realizadas** | **Desafíos** | **Puntaje** |
| Andrés Bravo | Ingeniero de software responsable de desarrollar la aplicación final | 10 | Gestionar el proceso de construcción de la aplicación y la conexión con el backend (API). | Hacer la aplicación más amigable al usuario, tanto en funcionalidad como en visualización. | 33.3 |
| Kevin Gámez | Ingeniero de datos | 10 | Creación del pipeline, velando por la calidad del proceso de automatización relacionado con la construcción del modelo analítico. | Exportar correctamente todo el preprocesamiento de los datos y el entrenamiento del modelo dentro de un solo archivo joblib. | 33.3 |
| Sara Plazas | Ingeniera de software responsable del diseño de la aplicación y resultados | 10 | Creación y diseño de la aplicación.  Generación del video con los resultados obtenidos. | Recordar elementos sobre el desarrollo web vistos en materias anteriores | 33.3 |

Después de una reflexión grupal se determinó que todos los integrantes del grupo aportaron al trabajo de su propia forma de manera equitativa y justa. Es por esto por lo que se tomó la decisión de repartir los puntos de forma igual entre los integrantes, es decir 1/3 para cada uno. Asimismo, creemos que es importante mantener las buenas prácticas que hemos adoptado como mantener una comunicación constante, entregar las tareas a tiempo y hacer una distribución razonable del trabajo. Con respecto a las oportunidades de mejora, se propone empezar los trabajos con un poco más de tiempo y así poder sacarles un mayor provecho a las herramientas y recursos del curso.