PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
Facultad de Ingeniería
IIC2613 Inteligencia Artificial
2023-2

---

# Tarea 4
# Procesamiento de Datos y clasificadores

Fecha de entrega: Jueves 16 de noviembre a las 23:59 hrs

---

## Aspectos generales

### Formato y plazo de entrega

El formato de entrega son archivos con extensión `.ipynb` con un PDF para las respuestas teóricas. El lugar de entrega es en el repositorio de la tarea, en la branch por defecto, hasta el jueves 16 de noviembre a las 23:59 hrs. Para crear tu repositorio, debes entrar en el enlace del anuncio de la tarea en Canvas. Por último, recuerda que los cupones de atraso son días **no hábiles** extra.

### Integridad Académica

Este curso se adhiere al Código de Honor establecido por la universidad, el cual tienes el deber de conocer como estudiante. Todo el trabajo hecho en esta tarea debe ser **totalmente individual**. La idea es que te des el tiempo de aprender estos conceptos fundamentales, tanto para el curso, como para tu formación profesional. Las dudas se deben hacer exclusivamente al cuerpo docente a través de las *issues* en GitHub.

Por otra parte, sabemos que estás utilizando material hecho por otras personas, por lo que es importante reconocerlo de la forma apropiada. Todo lo que obtengas de internet debes citarlo de forma correcta (ya sea en APA, ICONTEC o IEEE). Cualquier falta a la ética y/o a la integridad académica será sancionada con la reprobación del curso y los antecedentes serán entregados a la Dirección de Pregrado.

### Comentarios adicionales

El objetivo de esta tarea es que puedan utilizar algoritmos de aprendizaje de máquina para llevar a cabo tareas de clasificación sobre conjuntos de datos. Es fundamental que pongan énfasis en las justificaciones de sus respuestas, cuidando la redacción, ortografía; manteniendo el código ordenado y comentado. Aquellas respuestas que solo presenten resultados o código (sin contexto ni comentarios) no serán consideradas, mientras que tareas desordenadas pueden ser objeto de descuentos.

# 1. DCCine

En busca de un nuevo trabajo, te han contratado como crítico de películas. Sin embargo, no tienes experiencia y no sabes cómo dar opiniones sobre las filmaciones. Ante la necesidad de aprender, decidiste visitar la página de IMDb para aprender a realizar críticas basadas en las opiniones de diferentes usuarios. Encontraste una base de datos que contiene las reseñas de diversos usuarios y decidiste buscar ciertos patrones que te ayuden a crear tus propias críticas.


Figura 1: Plataforma IMDb

Como se mencionó en el apartado anterior, se te entregará el archivo `IMDb.csv` y, con base en eso, deberás realizar el procedimiento especificado en los siguientes puntos. Para esta parte, deberás entregar un archivo de Jupyter Notebook con las funciones y procedimientos realizados. Por otra parte, deberás entregar un archivo en formato PDF con las respuestas a las preguntas.

## 1.1. Lectura y estudio de los datos (0.3 ptos)

1. Utilizando *pandas*, lee la información contenida en el archivo `IMDB.csv`. Una vez que la información esté cargada, comenta sobre los elementos que contiene. Además, debes indicar cuál es el atributo que se utilizará como etiqueta para entrenar los modelos.

2. Utiliza las funciones describe() y count() para realizar un análisis inicial del DataFrame. Indica la cantidad de datos con los que se está trabajando, el tipo de datos presentes, la distribución de datos por clase y cualquier otra información que consideres relevante.

## 1.2. Preprocesamiento (1 pto)

En esta sección, deberás llevar a cabo un tratamiento de los datos para garantizar que los clasificadores puedan ser entrenados de manera adecuada. Es importante resaltar que todas las decisiones que tomes deberán estar debidamente justificadas.

1. Revisa si el DataFrame contiene valores nulos. En caso afirmativo, deberás tomar la decisión de eliminar las filas o columnas que contengan estos datos nulos, o si prefieres, reemplazar estos valores con algún otro valor. Si no existen valores nulos, indícalo. Finalmente, responde a la siguiente pregunta: ¿Por qué es importante eliminar los valores nulos? (0.2 ptos)

2. Considera la proporción de valoraciones positivas y negativas. Responde a las siguientes preguntas: (0.6 ptos)

- ¿Están balanceadas las clases?
- ¿Qué sucede si las clases están desbalanceadas?
- ¿Qué tratamientos se deben llevar a cabo para balancearlas?

Para este punto, debes investigar el procedimiento que se debe realizar si las clases están desbalanceadas e incluir la fuente bibliográfica de donde obtuviste la información. Finalmente, te recomendamos utilizar 5000 datos de cada clase, seleccionados de forma aleatoria.

3. Investiga el concepto de 'stopwords' y por qué es necesario eliminarlas del DataFrame. Después de realizar la investigación, lleva a cabo un procedimiento para excluir este tipo de palabras en el análisis. (0.2 ptos)

### 1.3. Vectorización (0.9 ptos)

En esta sección, es fundamental que transformes la información de las reseñas escritas en vectores numéricos que serán utilizados para entrenar los modelos. Para lograr esto, deberás hacer uso de Bag of Words, Tf-Idf y SBERT.

1. **Bag of Words:**
   Investiga la vectorización *Bag of Words*, explica su funcionamiento y aplícala sobre los datos procesados, luego almacena el resultado en una matriz llamada `X_bow`. (0.3 ptos)

2. **TF-IDF**
   Investiga la vectorización *TF-IDF*, explica su funcionamiento y aplícala sobre los datos procesados, luego almacena el resultado en una matriz llamada `X_tfidf`. (0.3 ptos)

3. **SBERT**
   Investiga la vectorización *SBERT*, explica su funcionamiento y aplícala a los datos procesados. Luego, almacena el resultado en una matriz llamada `X_sbert`. (0.3 ptos)

### 1.4. Utilización de Clasificadores (0.9 ptos)

En esta sección, deberás entrenar 9 modelos. Para hacerlo, se deben entrenar los clasificadores de árboles de decisión, Random Forest y SVM con las 3 matrices de vectores obtenidos anteriormente. Por ejemplo, tendrás 3 modelos de clasificador SVM: uno considerando la data Bag of Words, otro con la data de TF-IDF y, finalmente, otro con los datos de SBERT. Lo mismo se aplica a los clasificadores de árboles de decisión y Random Forest.

### 1.5. Explicabilidad usando Lime (0.45 ptos)

Investiga sobre la biblioteca LimeTextExplainer y aplícala a cada uno de los modelos entrenados para mostrar la explicación de los modelos, incluyendo las palabras clave.

### 1.6. Cálculo de métricas (0.9 ptos)

Genera una tabla comparativa en la que se indiquen las métricas de accuracy, precisión, recall y F1 score para cada uno de los modelos. Debes señalar cuál de los modelos obtuvo las mejores métricas y cuál de los clasificadores (árbol de decisión, Random Forest y SVM) presentó el mejor rendimiento. También menciona qué tipo de vectorización provocó que los modelos obtuvieran un mejor desempeño.

### 1.7. Análisis de desempeño (0.9 ptos)

Explica por qué un clasificador obtuvo un mejor desempeño que el otro y por qué cierta vectorización provocó que los clasificadores obtuvieran un mejor rendimiento. Para ello, basa tu respuesta en material bibliográfico.

### 1.8. Mejoramiento de los modelos (1.3 ptos)

1. Investiga cómo se puede obtener un modelo con mejor rendimiento y menciona todos los cambios pertinentes, como la adopción de un nuevo tipo de vectorización, la consideración de la cantidad total de datos a utilizar, la exploración de otros tipos de clasificadores, el uso de ensambles, entre otros. Deberás buscar información y citar las fuentes que utilices.

2. Implementa la solución propuesta a nivel de código, para ello te pediremos como mínimo que:

   - Utilizes un nuevo tipo de vectorización.
   - Utilizes algún clasificador (o una combinación de clasificadores) distinto a los entrenados previamente (no necesariamente debe ser un nuevo tipo de clasificador).
   - Compares mediante métricas de rendimiento el clasificador implementado contra los entrenados anteriormente.

### 1.9. Conclusiones (0.35 ptos)

¿Qué conclusiones puedes extraer de los diferentes modelos? Considera qué tipos de palabras son clave para determinar si una reseña termina siendo positiva o negativa.