

EVALUACIÓN	Obligatorio	GRUPO	FECHA
MATERIA	Machine Learning para Análisis de Datos		
CARRERA	Licenciatura en Sistemas		
CONDICIONES	<ul style="list-style-type: none"> - Puntaje máximo: 30 puntos - Puntaje mínimo: 0 puntos - Fecha de entrega: 10/12/2025 hasta las 21:00 horas en gestion.ort.edu.uy (max. 40Mb en formato zip, rar o pdf) 		
<p>Uso de material de apoyo y/o consulta</p> <p><u>Inteligencia Artificial Generativa</u></p> <ul style="list-style-type: none"> - Seguir las pautas de los docentes: Se deben seguir las instrucciones específicas de los docentes sobre cómo utilizar la IA en cada curso. - Citar correctamente las fuentes y usos de IA: Siempre que se utilice una herramienta de IA para generar contenido, se debe citar adecuadamente la fuente y la forma en que se utilizó. - Verificar el contenido generado por la IA: No todo el contenido generado por la IA es correcto o preciso. Es esencial que los estudiantes verifiquen la información antes de usarla. - Ser responsables con el uso de la IA: Conocer los riesgos y desafíos, como la creación de “alucinaciones”, los peligros para la privacidad, las cuestiones de propiedad intelectual, los sesgos inherentes y la producción de contenido falso. - En caso de existir dudas sobre la autoría, plagio o uso no atribuido de IAG, el docente tendrá la opción de convocar al equipo de obligatorio a una defensa específica e individual sobre el tema. <p>Defensa Fecha de defensa: 11/12/2025 Defensa en forma de pregunta en el 2do parcial.</p> <p>IMPORTANTE:</p> <ol style="list-style-type: none"> 1) Inscribirse 2) Formar grupos de hasta 2 personas del mismo dictado 3) Subir el trabajo a Gestión antes de la hora indicada (ver hoja al final del documento: “RECORDATORIO”) <p>Aquellos de ustedes que presenten alguna dificultad con su inscripción o tengan inconvenientes técnicos, por favor contactarse con el Coordinador de cursos o Coordinación adjunta antes de las 20:00h del día de la entrega, a través de los mails crosa@ort.edu.uy / posada_1@ort.edu.uy (matutino) / larrosa@ort.edu.uy (nocturno), o vía Ms Teams.</p>			

Predicción de Popularidad de Canciones en Spotify

Contexto del problema

Spotify, la plataforma líder de streaming musical, utiliza algoritmos complejos para determinar la popularidad de una canción, un factor clave para su visibilidad y éxito. Se les ha encomendado la tarea de desarrollar un modelo predictivo que pueda estimar la popularidad de una canción basándose en sus características musicales. Su objetivo es construir un **modelo de regresión** que, a partir de atributos como la energía, la bailabilidad y el tempo, prediga el puntaje de popularidad que una canción podría alcanzar en la plataforma.

Objetivos

Aplicar modelos de aprendizaje automático para abordar una tarea de regresión:

- Construir un modelo de regresión capaz de predecir la popularidad (*popularity*) de las canciones en Spotify.

Datos disponibles

El conjunto de datos contiene información detallada sobre las canciones disponibles en Spotify, algunas variables son:

Información adicional sobre las variables del dataset

- *artists* (categórica): nombre del/los artista/s.
- *album_name* (categórica): nombre del álbum al que pertenece la canción.
- *track_name* (texto): título de la canción.
- *popularity* (numérica): puntaje de popularidad de la canción (variable a predecir).
- *duration_ms* (numérica): duración de la canción en milisegundos.
- *explicit* (booleana): indica si la canción contiene contenido explícito.
- *danceability* (numérica): describe qué tan adecuada es una canción para bailar.
- *energy* (numérica): medida de intensidad y actividad.
- *key* (numérica): clave musical de la canción.
- *loudness* (numérica): volumen general de la canción en decibeles.
- *mode* (categórica): indica si la canción está en tono mayor o menor.
- *tempo* (numérica): velocidad de la canción en beats por minuto (BPM).
- *time_signature* (numérica): estimación del compás de la canción.
- *track_genre* (categórica): género al que pertenece la canción.

Competencia Kaggle

Como parte fundamental del trabajo, se deberá participar en una competencia privada en la plataforma Kaggle. El objetivo de la competencia es predecir la variable `popularity` sobre un conjunto de canciones de test para el cual no se conoce su valor real.

Se deberá realizar al menos un envío con las predicciones generadas por el modelo entrenado, lo cual es de carácter obligatorio para aprobar el trabajo. El conjunto de test no contiene la variable `popularity`, y será evaluado por la plataforma utilizando la métrica ***Root Mean Squared Error (RMSE)***. El objetivo es obtener un valor de RMSE lo más bajo posible.

Aclaración: La posición obtenida en el ranking de la competencia no será un factor condenatorio en la nota del obligatorio. El foco está en el proceso, la justificación de las decisiones y el aprendizaje.

Link a la competición de Kaggle

Formato de entrega y tareas a desarrollar

Se deberá entregar un documento (en formato pdf) que contenga un **informe/reporte** de los resultados obtenidos, detallando las diferentes experimentaciones realizadas. El documento puede estructurarse, por ejemplo, en secciones como Metodología, Resultados y Conclusiones.

Es importante destacar que **la performance del modelo no es el único aspecto relevante de la entrega**. La calidad de la documentación, la justificación de las decisiones tomadas y las reflexiones obtenidas a partir de los experimentos realizados son elementos fundamentales del trabajo. Se espera que los estudiantes analicen críticamente los resultados y argumenten con claridad sus conclusiones, más allá del resultado obtenido en la competencia.

Además del informe, **deberá entregarse todo el código desarrollado durante la experimentación**, incluyendo notebooks o scripts utilizados para el preprocessamiento, entrenamiento de modelos, evaluación, comparación y generación de predicciones.

Se deberán desarrollar las siguientes tareas que deberán estar incluídas en el informe:

1. **Preprocesamiento:** Definir y justificar el tratamiento inicial realizado a los datos. Es crucial que cada decisión tomada sea explicada en detalle, incluyendo el análisis de su impacto en el modelo. Las tareas mínimas a realizar son:
 - **Gestión de valores faltantes:** Analizar la distribución de los valores nulos y aplicar una estrategia justificada de imputación o eliminación.
 - **Gestión de outliers:** Identificar y tratar los valores atípicos en las variables numéricas relevantes, explicando la razón detrás de la elección del método.
 - **Transformación de variables:** Aplicar transformaciones (ej. *logarítmica*) a las variables que crean necesarias.

2. Feature Engineering: Desarrollar un proceso de creación, modificación y selección de variables que mejore el rendimiento del modelo. Es fundamental justificar cada decisión. Se espera que realicen, al menos:

- **Manejo de variables categóricas:** Transformar las variables categóricas (ej., artists, album_name) en un formato adecuado para el modelado, considerando el impacto de la alta cardinalidad.
- **Creación de nuevas características:** A partir de las variables existentes, crear al menos una nueva característica que aporte valor predictivo al modelo.
- **Selección de características:** Justificar la eliminación de columnas que consideren irrelevantes o que no aportan información significativa.
- **Gestión de *data leakage*:** Asegurar que todos los procesos de creación de características y escalado de datos se realicen de manera correcta para evitar la fuga de información del conjunto de prueba.

3. Modelos de Regresión: Se requiere la implementación de los siguientes modelos de regresión, comparando su desempeño en la tarea de predicción.

- Regresión Lineal.
- Árbol de Decisión para Regresión.
- Random Forest Regressor.
- Gradient Boosting Regressor.
- Redes Neuronales.

4. Evaluación y selección de modelos:

- Detallar el proceso de evaluación (validación) y selección de los diversos modelos de regresión, incluyendo la búsqueda de hiperparámetros. Utilizar **alguna** de las técnicas vistas en el curso (Holdout, Repeated Holdout o Cross-Validation).
- Para cada experimento, registrar las métricas relevantes al problema de regresión: **Root Mean Squared Error (RMSE)**, **Mean Squared Error (MSE)**, **R² Score** y **Mean Absolute Error (MAE)**.
- Resumir los resultados de los experimentos en tablas que permitan recopilar la información de cada modelo evaluado y sus respectivas métricas. Se recomienda incluir una tabla por cada algoritmo que resuma el desempeño del mismo para distintos valores de los hiperparámetros, así como una tabla final que resuma el desempeño de los mejores modelos.
- Participar de la competencia de Kaggle. Incluir en el informe el desempeño de aquellos modelos sometidos a la competencia.

5. Modelo final: se deberá seleccionar un modelo final, justificando claramente la elección en función de las métricas obtenidas, la robustez del modelo y las observaciones realizadas durante la experimentación.

El informe deberá incluir:

- Una explicación clara de por qué se eligió ese modelo como el más adecuado.
- Una breve descripción de los hiperparámetros seleccionados y su impacto.
- Las predicciones finales generadas para el conjunto de test.

El **código completo utilizado para entrenar, validar y generar predicciones con el modelo final** deberá ser entregado en formato Jupyter notebook (Python), **con las celdas ya ejecutadas**.

RECORDATORIO: IMPORTANTE PARA LA ENTREGA

• Obligatorios

La entrega de los obligatorios será en formato digital online, a excepción de algunas materias que se entregarán en Bedelía y en ese caso recibirá información específica en el dictado de la misma.

Los principales aspectos a destacar sobre la **entrega online de obligatorios** son:

1. Ingresá al sistema de Gestión.
2. En el menú, seleccioná el ítem “Evaluaciones” y la instancia de evaluación correspondiente, que figura bajo el título “Inscripto”.
3. Para iniciar la entrega hacé clic en el ícono: 
4. Ingresá el número de estudiante de cada uno de los integrantes y hacé clic en “Agregar”. El sistema confirmará que los integrantes estén inscriptos al obligatorio y, de ser así, mostrará el nombre y la fotografía de cada uno de ellos. Una vez agregados todos los integrantes, hacé clic en “Crear equipo”.
Cualquier integrante podrá:
 - **Modificar la integración del equipo.**
 - **Subir el archivo de la entrega.**
5. Seleccioná el archivo que deseás entregar. Verificá el nombre del archivo que aparecerá en la pantalla y hacé clic en “Subir” para iniciar la entrega. Cada equipo (hasta 2 estudiantes) debe entregar **un único archivo en formato zip o rar** (los documentos de texto deben ser pdf, y deben ir dentro del zip o rar). El archivo a subir debe tener **un tamaño máximo de 40mb**.
Cuando el archivo quede subido, se mostrará el nombre generado por el sistema (1), el tamaño y la fecha en que fue subido.
6. El sistema enviará un e-mail a todos los integrantes del equipo informando los detalles del archivo entregado y confirmando que la entrega fue realizada correctamente.
7. Podés cerrar la pestaña de entrega y continuar utilizando Gestión o salir del sistema.
8. La **hora tope para subir el archivo será las 21:00** del día fijado para la entrega.
9. La entrega se podrá realizar desde cualquier lugar (ej. hogar del estudiante, laboratorios de la Universidad, etc.).
10. Aquellos de ustedes que presenten alguna dificultad con su inscripción o tengan inconvenientes técnicos, por favor contactarse con el Coordinador de cursos o Coordinación adjunta antes de las 20:00h del día de la entrega, a través de los mails crosa@ort.edu.uy / posada_1@ort.edu.uy (matutino) / larrosa@ort.edu.uy (nocturno), o vía Ms Teams.