



Primer modelo de Machine Learning

Vas a poder entregar desde el
25 de Marzo a las 21:00hs.

 Individual

La realización y entrega del Proyecto es individual

Un Notebook de Jupyter con la resolución de la consigna. El Notebook debe poder ejecutarse sin errores.

Para realizar el proyecto debes usar como referencia el contenido de las bitácoras, de los notebooks trabajados y las presentaciones vistas en clase. También te será de mucha utilidad consultar la documentación de las librerías de Python que trabajamos en clase, consultar comunidades online como Stack Overflow y, por supuesto, buscar en la web (*googlear*).

Utiliza los siguientes recursos para realizar el proyecto:

- Notebook de Jupyter (*DSProyecto01.ipynb*)
- Dataset Properati (*DSProyecto01DatosProperati.csv*)

PARTE 1 – Pensando como um/a Data Scientist

En esta sección, te planteamos una problemática y deberás responder la siguiente pregunta:

¿Qué datos crees que te ayudarían a trabajar en el problema? ¿Por qué?

Importante: NO deberás buscar esos datos, solamente justificar qué información crees que te ayudaría a resolver la problemática planteada.

PARTE 2 – Análisis Exploratorio de Datos

En esta sección, te proveeremos de un Dataset relacionado con la problemática planteada para que realices un **Análisis Exploratorio de Datos**. Deberás responder las preguntas típicamente asociadas a este análisis para obtener una primera descripción del Dataset.

PARTE 3 – Primer Modelo de Machine Learning

Una vez explorado el dataset, deberás utilizar herramientas de Machine Learning para predecir la variable de interés. Para ello, tendrás que:

1. Elegir la métrica que utilizarás para evaluar las predicciones.
2. Generar tres modelos: un modelo Benchmark, un modelo basado en Vecino más Cercanos (Modelo 1) y otro basado en Árboles de Decisión (Modelo 2). Cada uno de estos modelos deberá ser correctamente entrenado y evaluado siguiendo un flujo de trabajo típico de Machine Learning (train/test split).
3. Optimizar el desempeño de los Modelos 1 y 2 optimizando el número de vecinos y la profundidad del árbol, respectivamente.
4. Comparar la performance de los Modelos 1 y 2 frente al modelo Benchmark y entre sí para decidir cuál modelo tiene mejor desempeño.

Sugerencias para desarrollar el proyecto:

- Debes resolver los **mínimos entregables** indicados en la consigna, pero te invitamos a que siempre profundices más.
- La resolución del proyecto te puede enfrentar a pequeños desafíos que no trabajamos durante los encuentros. Es importante que desarrolles la capacidad de resolverlos. Para ello, consulta las bitácoras, las referencias, los notebooks vistos en clase, la documentación de la librería y, sobre todo, *googlea*.
- Recuerda que un notebook es un informe, por lo que debes ir explicando lo que haces a medida que resuelves las consignas. Es importante que quien te lo lea entienda el flujo de trabajo, qué quisiste hacer. Recuerda, simple y sencillo es una combinación ganadora.

Antes de subir tu proyecto a la plataforma Académica para que sea evaluado, verifica que el Notebook se ejecute sin errores. Además, asegúrate de cumplir con las siguientes condiciones (son las que los/as evaluadores/as tendrán en

PARTE 1 – Pensando como un/a Data Scientist

1. Debes justificar por qué creés que los datos que elegiste ayudan a resolver la problemática planteada.

PARTE 3 – Análisis Exploratorio de Datos

1. Debes responder las preguntas típicas de un Análisis Exploratorio de Datos.
2. Debes utilizar histogramas para representar la distribución de variables numéricas, gráficos de barras para variables categóricas y diagramas de dispersión para visualizar la relación entre dos variables numéricas. Si no exploras alguna variable (columna), deberás justificar por qué.
3. Los gráficos deben estar correctamente presentados, con título y etiquetas en cada eje. Cada gráfico debe ir acompañado de una breve descripción. Si existen valores atípicos que distorsionan la visualización de un gráfico, el límite de cada eje debe estar ajustado para que esto no suceda o el dataset debe ser filtrado.

PARTE 3 – Primer Modelo de Machine Learning

1. La elección de la métrica de evaluación debe estar correctamente justificada.
2. Debes elegir correctamente las variables predictoras y la variable a predecir.
3. Realiza un train/test split de los datos.
4. El modelo benchmark debe estar justificado y evaluado de forma similar a los modelos generados.
5. Debes justificar cuál modelo elegirías para utilizar.
6. Debes ser **crítico/a** con la metodología utilizada. ¿Qué mejorarías?