GUIA DE ACTIVIDADES HACIA EL PROYECTO FINAL

Data Science II: Machine Learning para la Ciencia de Datos

CODERHOUSE

¡Bienvenidas y bienvenidos!

Qué bueno encontrarlos/as en este espacio el cual hemos creado para que puedan conseguir en un mismo lugar, de manera rápida y ágil, todas las actividades prácticas entregables que plantea el curso.

A continuación presentamos el sistema de entregas de los cursos de Coder. Luego, en un tablero, podrán ver **las clases establecidas en el programa** marcando con el ícono correspondiente las clases que sí tienen entregables (incluido el proyecto final).

De esta forma podrás tener un pantallazo del cronograma de clases y las actividades prácticas que deberás completar.

CODERHOUSE

Sistema de entregas

actividad prácticas entregables

Tienen una vigencia de 7 días, es decir, a partir de la fecha (de la clase) en la que se lanza la actividad práctica, empezarán a correr 7 días continuos, para que puedas cargar tu actividad práctica en la plataforma.

Pre-entrega del PF

También tiene una vigencia o duración de 7 días antes de que el botón de “entrega” se deshabilite. Por este motivo te recomendamos estar al día con todas las actividades planteadas.

Proyecto Final

A diferencia de los anteriores puntos, cuenta con un lapso de 20 días continuos luego de finalizada la última clase. Posterior a ese tiempo, el botón de la entrega quedará inhabilitado y no será posible entregarlo o recibirlo por otros medios.

GRILLA DE ENTREGAS – MACHINE LEARNING PARA LA CIENCIA DE DATOS

Clases	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Actividad práctica															
Proyecto final															

Clases	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	27	28	29	30
Actividad práctica															
Proyecto final															

CODERHOUSE

CLASE 2

Data Acquisition

Descarga de datos desde APIs públicas

Crearás un notebook donde se seleccionará una API de interés, luego crearás una API key y finalmente extraerás la información para ser almacenada en un DataFrame

CODERHOUSE

Descarga de datos desde APIs públicas

Consigna

- ✓ Buscar información en APIs públicas (i.e Twitter, NewsAPI, Spotify, Google Apis, etc).
- ✓ Extraer datos e importarlos a un dataframe realizando una exploración simple.

Aspectos a incluir

- ✓ Notebook donde se detallen todos los pasos seguidos

Ejemplo

- ✓ Ejemplo APIS,

Formato

- ✓ Se debe entregar un jupyter notebook con el nombre "**actividad práctica_APIS_+Nombre_+Apellido.ipynb**".

Sugerencias

- ✓ No compartir sus tokens personales
- ✓ Comprender primero el funcionamiento de las APIs a detalle para después utilizarla
- ✓ La limpieza de datos en APIS no es fácil
- ✓ Tratar de obtener datos principalmente numéricos (al menos 20 columnas y 10000 filas.
- ✓ Si realizaron el curso anterior, Se sugiere que estos datos complementen el dataset elegido en el práctica "Elección de potenciales Datasets e importe con la librería Pandas"

CLASE 8

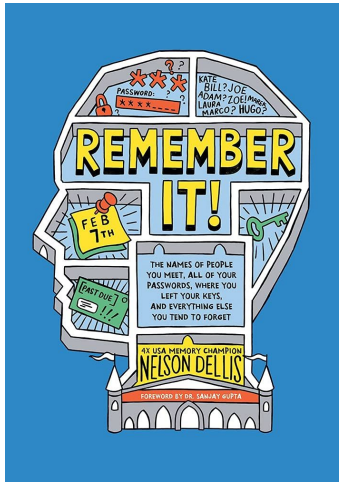
Data Wrangling

Data Wrangling

Continuaremos hablando sobre lo trabajado en la práctica anterior.. Crearás un notebook donde se desarrollará la limpieza de los datos elegidos para tu proyecto final, deberás tener en cuenta técnicas vistas en clase para el tratamiento de valores duplicados, nulos y outliers con su respectiva justificación.

CODERHOUSE

Recordemos...



- Exploramos diversas APIs
- Extrajimos los datos en un dataframe
- Realizamos una exploración simple
- Obtención de Insights preliminares

CODERHOUSE

Data Wrangling

Consigna

- ✓ Iniciar el proceso de limpieza y exploración de datos.
- ✓ Explorar los tipos de las columnas y modificar según corresponda.
- ✓ Validar la presencia de valores perdidos y sugerir alguna solución.
- ✓ Validar la presencia de valores outliers y sugerir alguna solución.

Aspectos a incluir

- ✓ Notebook con código y estructura eficiente

Ejemplo

- ✓ Data Wrangling

Formato

- ✓ Se espera un notebook en formato .ipynb. Dicho notebook debe tener el siguiente nombre "**Data_Wrangling+Apellido.ipynb**".

Sugerencias

- ✓ Utilizar las herramientas vistas en el curso
- ✓ Manejo de duplicados nulos y análisis exploratorio

Explicación dla actividad práctica

- ✓ ¡Click aquí!

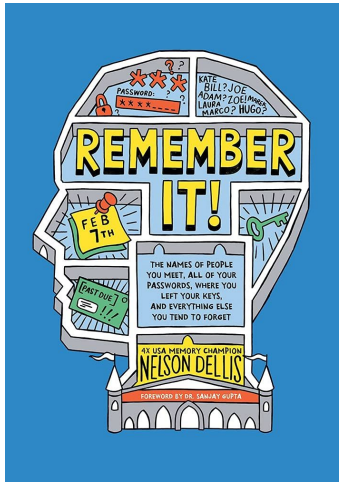
CLASE 12

Visualización efectiva y Data Storytelling

Data Storytelling

Continuaremos hablando sobre lo trabajado en la práctica de "**Data Wrangling"**. Crearás un notebook donde se desarrolle una narrativa que permita dar respuesta a las preguntas/hipótesis formuladas para el proyecto final.

Recordemos...



- Extrajimos datos de interés
- Comenzamos el proceso de limpieza y estructuración
- Desarrollamos algunas hipótesis

- Data Wrangling/Munging

CODERHOUSE

Data Storytelling

Consigna

- ✓ Generar preguntas de interés o hipótesis de interés sobre el dataset elegido para el proyecto final.
- ✓ Iniciar el proceso de Data Storytelling avanzando sobre las posibles respuestas a aquellas.
- ✓ Mejorar y hacer nuevas gráficas según las técnicas aprendidas de formato.

Aspectos a incluir

- ✓ Notebook con código y estructura eficiente

Ejemplo

- ✓ Data StoryTelling

Formato

- ✓ Se espera un notebook en formato .ipynb. Dicho notebook debe tener el siguiente nombre "**Data_StoryTelling+Apellido.ipynb**".

Sugerencias

- ✓ Utilizar las herramientas vistas en el curso
- ✓ Manejo de duplicados nulos y análisis exploratorio
- ✓ Comenzar por preguntas de alto nivel y luego más específicas

Explicación dla actividad práctica

- ✓ ¡Click aquí!

CLASE 17

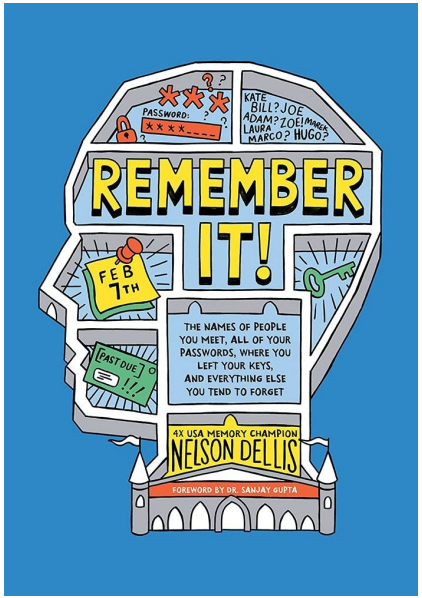
Workshop: Revisión de pares

Primera preentrega de tu Proyecto final

Entrenarás y optimizarás diversos modelos de Machine Learning para resolver una problemática específica, detectada en la instancia de entrega anterior. El objetivo es que puedan utilizar modelos de Machine Learning para resolver el problema de una industria o negocio.

CODERHOUSE

Recordemos...



Generamos visualizaciones

Obtuvimos insights

Desarrollamos una narrativa correcta

Responder las preguntas problema

CODERHOUSE

Obtención de insights a partir de visualizaciones

Objetivos generales

- ✓ Obtener datos de diversas fuentes como APIs o Bases de datos públicas para luego analizarlos mediante el lenguaje Python con el fin de contestar una pregunta de interés para una industria, negocio o proyecto personal. Se deberán utilizar datasets complejos implementando técnicas avanzadas para la limpieza y adquisición de datos

Objetivos específicos

- ✓ Estructurar un problema en función de múltiples pero simples preguntas/hipótesis a responder
- ✓ Importar datos crudos de APIs o bases de datos usando Python
- ✓ Limpiar y transformar los datos para permitir un posterior análisis
- ✓ Contar una historia mediante el análisis exploratorio de datos

Obtención de insights a partir de visualizaciones

Requisitos base

✓ Un notebook (Colab o Jupyter) que debe contener:

1. **Abstracto con motivación y audiencia:** Descripción de alto nivel de lo que motiva a analizar los datos elegidos y que audiencia se podrá beneficiar de este análisis
2. **Preguntas/hipótesis que queremos responder:** Lista de preguntas que se busca responder mediante el análisis de datos. Bloques de código donde se importan los datos desde una API o base de datos pública y los guarda en un archivo local csv o json. El estudiante puede luego de descargar los datos, comentar este bloque de código
3. **Análisis exploratorio de datos (EDA):** Análisis descriptivo de los datos mediante visualizaciones y herramientas estadísticas

Obtención de insights a partir de visualizaciones

Requisitos base

✓ Una presentación (PDF; PowerPoint o Google Slides) que debe contener

1. **Abstracto con motivación y audiencia:** Descripción de alto nivel de lo que motiva a analizar los datos elegidos y que audiencia se podrá beneficiar de este análisis
2. **Resumen de metadata:** resumen de los datos a ser analizados es decir, número de filas/columnas, tipos de variables, etc
3. **Preguntas hipótesis que queremos responder:** Lista de preguntas que se busca responder mediante el análisis de datos
4. **Visualizaciones ejecutivas que responden nuestras preguntas:** utilización de gráficos que responden las preguntas de interés de nuestro proyecto.
5. **Insights:** resumen de hallazgos del proyecto. Aquí consolidamos las respuestas a las preguntas/hipótesis que fuimos contestando con las visualizaciones

Obtención de insights a partir de visualizaciones

Sugerencias

Es conveniente retomar el dataset trabajado en la primera pre entrega y enriquecerlo (e.g joins, y creación de nuevas columnas) con información proveniente de APIs públicas siempre que se pueda con el fin de practicar las nuevas habilidades adquiridas. Se recomienda retomar la metodología de trabajo y reutilizar algoritmos ya entrenados, de ser necesario.

Requisitos extra

- ✓ Subir el proyecto a Github

Obtención de insights a partir de visualizaciones

Dont's

- ✓ Utilizar jerga demasiado técnica en la presentación (recordar que la audiencia de la misma son roles ejecutivos)
- ✓ Sobrecargar las diapositivas
- ✓ Realizar una presentación con más de 12 slides de extensión

Modelo de Proyecto final

- ✓ Proyecto final (Notebook) (Se debe abrir con Google Collaboratory o Jupyter Notebook)
- ✓ Ejemplo Presentación

Explicación dla actividad práctica

- ✓ ¡Click aquí!

CLASE 21

Algoritmos de clasificación y regresión

Entrenando un algoritmo de Machine Learning

Deberás entregar el avance de tu proyecto final. Continuaremos hablando sobre lo trabajado en la **segunda pre entrega del proyecto final**. Crearás un notebook donde trabajarás sobre los datos elegidos en la primera y segunda pre entrega del proyecto final. Posteriormente, realizarás las etapas de: i) Encoding, ii) Ingeniería de atributos y iii) Entrenamiento de un modelo de Machine Learning Supervisado (Clasificación o Regresión) o no supervisado dependiendo de la pregunta problema.

CODERHOUSE

ACTIVIDAD PARA EL PROYECTO FINAL

Entrenando un algoritmo de Machine Learning

Consigna

- ✓ Utilizar su dataset para resolver problemas de clasificación o regresión.
- ✓ Realizar tareas de preprocesamiento, por ejemplo encoding.
- ✓ Ejecutar el proceso de feature engineering para optimizar el tamaño del dataset sin perder poder predictivo.
- ✓ Separar el dataset en train y test.
- ✓ Entrenar su primer modelo de Machine Learning (Clasificación o Regresión) según corresponda.

Aspectos a incluir

- ✓ Notebook donde se detallen todos los pasos seguidos

Ejemplo

- ✓ Ejemplo actividad práctica Entrenamiento

Formato

- ✓ Se debe entregar un jupyter notebook con el nombre "**actividad práctica_AlgoritmoML_MVP_+Nombre_ +Apellido.ipynb**".

Sugerencias

- ✓ Se pueden utilizar fuentes de datos conocidas en sitios como Kaggle o UCI
- ✓ Se recomienda elegir datasets curados para que la mayor parte del tiempo se utilice para el entrenamiento de modelos y no en limpieza de datos.

Explicación dla actividad práctica

- ✓ ¡Click aquí!

CODERHOUSE

CLASE 27

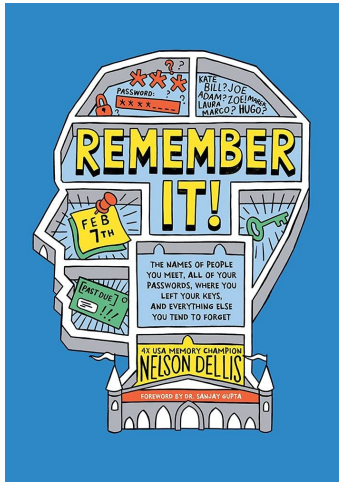
Validación de modelos - métricas

CODERHOUSE

Ingeniería de atributos y selección de variables

Crearás un notebook donde se terminará el proceso de Feature Engineering dla actividad práctica anterior, se busca que se puedan crear nuevas variables sintéticas que ayuden a mejorar el desempeño de los modelos de Machine Learning. Finalmente, deberás realizar un PCA sobre todas las variables utilizadas con el fin de determinar el peso relativo de cada variable en los modelos.

Recordemos...

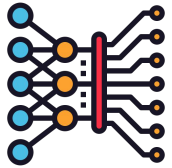


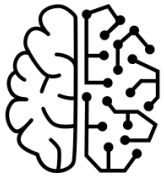
Terminamos de realizar el proceso de Feature Engineering

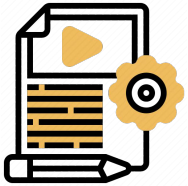
Ampliamos el número de variables en el MVP

Realizamos una segunda ronda de entrenamiento con más variables

Evaluamos al algoritmo







CODERHOUSE

Ingeniería de atributos y selección de variables

Consigna

- ✓ Crear variables sintéticas adicionales que permitan mejorar el desempeño del modelo de ML..
- ✓ Testear distintos modelos.
- ✓ Determinar si alguno presenta over o under fitting y relacionar la conclusión con el trade-off entre bias y variance.
- ✓ Realizar PCA sobre las variables usadas y explorar las cargas de los 2 primeros componentes, identificar las variables más relevantes

Aspectos a incluir

- ✓ Notebook donde se detallen todos los pasos seguidos

Ejemplo

- ✓ Feature Selection (Filter Method),

Formato

- ✓ Se debe entregar un jupyter notebook con el nombre "**actividad práctica_FeatureSelection_+Nombre_ +Apellido.ipynb**".

Sugerencias

- ✓ Se recomienda realizar el PCA con el fin de obtener las variables sintéticas y reducir el número de inputs con el fin de mejorar el desempeño de los modelos de ML elegidos
- ✓ Dedicar un buen tiempo a la explicación de la metodología usada

Explicación dla actividad práctica

- ✓ ¡Click aquí!

CODERHOUSE

CLASE 30

Datathon

Entrenamiento y optimización de Modelos de Machine Learning

Deberás entregar tu **Proyecto Final.** Entrenarás y optimizarás diversos modelos de machine learning para resolver una problemática específica, detectada en la instancia de entrega anterior. El objetivo es que puedas utilizar modelos de ML para resolver el problema de una industria o negocio

CODERHOUSE

Entrenamiento y optimización de modelos de Machine Learning

Objetivos generales

- ✓ Utilizar modelos de Machine Learning para resolver un problema de una industria o negocio

Objetivos específicos

- ✓ Retomar el trabajo realizado en la segunda pre entrega, sumando el trabajo con Machine Learning
- ✓ Modelar la situación como un problema de Machine Learning
- ✓ Entrenar modelos de Machine Learning
- ✓ Realizar ingeniería de atributos y normalización/estandarización de variables
- ✓ Seleccionar el modelo con mejor performance

Entrenamiento y optimización de modelos de Machine Learning

Requisitos base

✓ Un Notebook (Colab o Jupyter) que debe contener:

1. **Abstracto con motivación y audiencia:** Descripción de alto nivel de lo que motiva a analizar los datos elegidos y audiencia que se podría beneficiar de este análisis.
2. **Preguntas/Problema que buscamos resolver:** Si bien puede haber más de una problemática a resolver, la problemática principal debe encuadrarse como un problema de clasificación o regresión.
3. **Breve Análisis Exploratorio de Datos (EDA):** Análisis descriptivo de los datos mediante visualizaciones y herramientas estadísticas, análisis de valores faltantes.
4. **Ingeniería de atributos:** Creación de nuevas variables, transformación de variables existentes (i.e normalización de variables, encoding, etc.)
5. **Entrenamiento y Testeo:** Entrenamiento y testeo de al menos 2 modelos distintos de Machine Learning utilizando algún método de validación cruzada.
6. **Optimización:** Utilizar alguna técnica de optimización de hiperparámetros (e.g gridsearch, randomizedsearch, etc.)
7. **Selección de modelos:** utilizar las métricas apropiadas para la selección del mejor modelo (e.g AUC, MSE, etc.)

Entrenamiento y optimización de modelos de Machine Learning

Piezas sugeridas

✓ **Librerías:**

numpy - pandas - matplotlib - sklearn - xgboost - shap

✓ **Claridad de código:**

Estructura - Markdown - Comentarios

Modelo de Proyecto final

✓ Ensemble Models

Explicación dla actividad práctica

✓ ¡Click aquí!

Para finalizar...

Llegar hasta aquí se traduce en esfuerzo, sudor, dedicación, trabajo, lágrimas, coraje y más, así que **FELICITACIONES** por haber dado el 100% de ti en toda la cursada.

Esto no termina acá porque recuerden que **cuentan con 10 días corridos para la entrega del Proyecto Final**, el cual estamos seguros de que quedará impecable.

Celebramos que ya estén coderizados y que vayan a donde vayan, ¡tendrán éxito! Que sigan alcanzando todo lo que se propongan 🚀

CODERHOUSE