aprendé programando

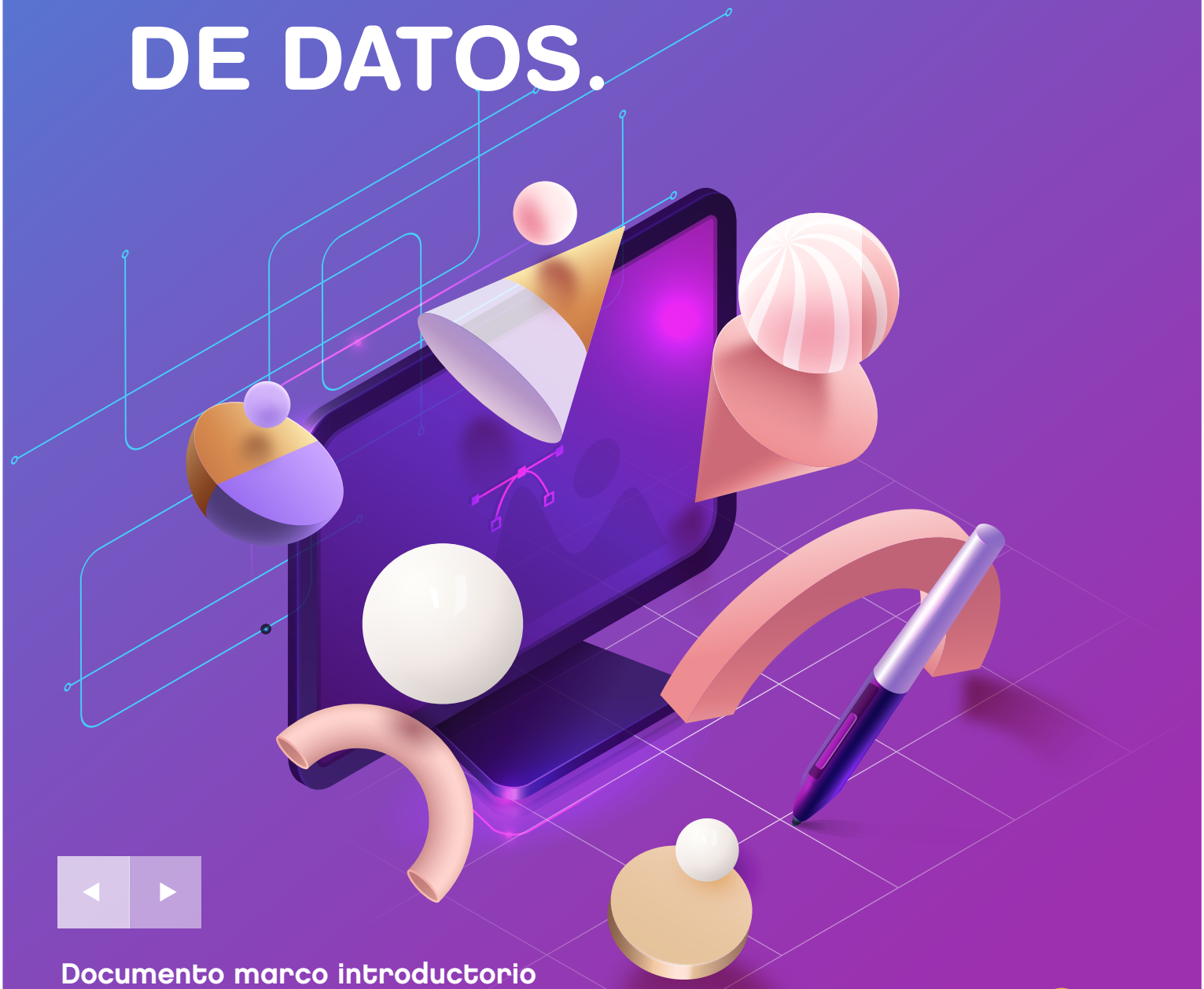
GyEE Gestión y Evaluación Educativa

UTN F r a

Módulo IV

El mundo y sus problemas a través de los lentes de un/a científico/a

CIENCIAS DE DATOS.



Documento marco introductorio

Buenos Aires Ciudad

Vamos Buenos Aires

Autoridades

Jefe de Gobierno
Horacio Rodríguez Larreta

Ministra de Educación
María Soledad Acuña

Jefe de Gabinete
Luis Bullrich

Subsecretario de Carrera Docente
y Formación Técnica Profesional
Manuel Vidal

Subsecretaria de Coordinación
Pedagógica y Equidad Educativa
María Lucía Feced Abal

Subsecretaria Agencia de Aprendizaje
a lo Largo de la Vida
Eugenia Cortona

Subsecretario de Gestión Económico
Financiera y Administración de Recursos
Sebastián Tomaghelli

Subsecretario de Tecnología
Educativa y Sustentabilidad
Santiago Andrés

El mundo y sus problemas
a través de los lentes
de un/a científico/a

Módulo IV

CIENCIAS DE DATOS.

- PROYECTO INTEGRADOR FINAL.
  NUEVOS DESAFÍOS.

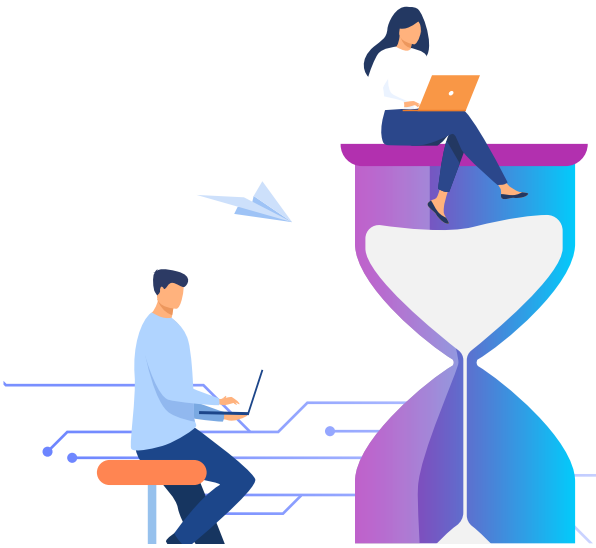
Orientaciones para los/las mentores/as

Índice

Índice

01. Sobre el módulo

En este módulo se trabajará en conocer modelos y herramientas de aprendizaje automático más complejas que las vistas hasta el momento.
El objetivo principal es que los/as estudiantes puedan terminar los Proyectos entrenando modelos de Aprendizaje Automáticode acuerdo al tema elegido y con el dataset que estuvieron trabajando el Módulo anterior. Al finalizar deben realizar una Presentación contando lo que hicieron a una audiencia no-técnica.
A su vez cada encuentro tendrá un espacio para introducir temas desde un lugar teórico (y a veces práctico) que tocan la Ciencia de Datos pero que quedarán para más adelante como ***CPU, GPU, TPU: requerimientos técnicos detrás de los algoritmos, Cloud computing y lo definido por el/la mentor/a de acuerdo a la comisión***
Estos bloques se llamarán "Una ventana para mirar" y tienen como objetivo realizar un primer acercamiento conceptual sobre cada una de estas complejas temáticas, para despertar intereses y seguir ampliando el conocimiento de los/las estudiantes en el mundo de la ciencia de datos.



Las últimas clases de este Módulo serán definidas por el/la mentor/a de acuerdo a las necesidades de cada comisión.
Se disponibilizarán algunos posibles temas para conocer, pero se intentará que cada mentor/a defina de acuerdo a los interes de los/as estudiantes. En el caso que sea necesario también se utilizará ese espacio para avanzar en los Proyectos o repasar algún tema que hayan quedado dudas.

Semana 1

Objetivos

- Conocer ensambles de Modelos: Random Forest
- Comprender mejor las unidades de procesamiento a la hora de trabajar con algoritmos y Aprendizaje Automático.
- Avanzar en los Proyectos

Actividad 1
Conociendo ensambles: Random Forest

Encuentro sincrónico

"Random Forest es uno de los algoritmos más utilizados en la técnica de regresión, y es un algoritmo de Machine Learning muy flexible y fácil de usar, incluso sin ajuste de hiperparámetros. Además, este algoritmo se usa ampliamente debido a su simplicidad y al hecho de que puede usarse tanto para tareas de regresión como de clasificación. El Forest que construye es un conjunto de árboles de decisión, la mayoría de las veces entrenados con el método de bagging" .

Seguimos sumando modelos posibles para aplicar en los Proyectos, en este caso presentaremos con una PPT el concepto de ensambles y el algoritmo Random Forest. Se verán sus aplicaciones y se correrá un simple problema de clasificación a modo de ejemplo con un dataset de *vidrio.*

Actividad 2
Comenzamos a esbozar el código que nos permitirá correr los modelos

Como en todas las clases, se dará un espacio en los encuentros para realizar el seguimiento de los Proyectos, los/as estudiantes deberán ya estar comenzando en la división de los datos y pensando qué modelos entrenar. El/La mentor/a deberá hacer un seguimiento de cada grupo para ver si hay dudas o necesidad de ayuda con estos temas.

Actividad 3
Una ventana para mirar: CPU, GPU, TPU: requerimientos técnicos detrás de los algoritmos

Con el tiempo restante de la clase introduciremos el tema de qué son las unidades de procesamiento y cuales son sus diferencias. La idea es dar un pantallazo general para que luego ellos y ellas investiguen y suban sus investigaciones al espacio asincrónico.

Encuentro asincrónico

Se este destinará un espacio de seguimiento a los proyectos. Cada grupo deberá avanzar en la división de datos y comenzando a definir modelos para entrenar.

Materiales necesarios:

> PPT *Random Forest*
> Notebook *Random Forest*
> Dataset *Glass*
> PPT CPU, GPU y TPU

Semana 2

Objetivos

- Aprender sobre *Grid-Search*
- Conocer el mundo de Cloud Computing y sus beneficios
- Avanzar en los Proyectos

Encuentro sincrónico

Actividad 1
Probando hiperparámetros

En este encuentro conocermos una herramienta para mejorar la performance de los modelos a través de la búsqueda de hiperparámetros. Para esto se conocerá el concepto de *Cross.-validation* con una PPT y luego se mostrará con una Notebook un ejemplo de Random Forest con el Dataset de Titanic extraído de Kaggle **[1]** que ya tiene realizada una limpieza.

Esto nos servirá para repasar Random Forest y utilizar este ejemplo para entrenar modelos con distintos hiperparámetros. Hay que tener en cuenta que está Notebook tal vez tarde un poco más en ejecutarse.
Esto a su vez serivirá para mostrar a los/as estudiantes los tiempos de procesamiento que pueden demorar dependiendo lo que estemos haciendo y el dataset que estemos usando.
Esta instancia deberá tener en cuenta el poder de procesamiento de las computadoras y cada equipo elegirá si todo el mundo va a correr los modelos en sus computadoras o si algunas no cuentan con los requisitos técnicos mínimos para poder hacerlo.

Actividad 2
Seguimiento de Proyectos

El espacio de seguimiento de Proyectos, el/la mentor/a deberá observar los avances y acompañar en caso que sea necesario. Para este momento sería bueno que ya hayan podido entrenar un modelo y estén pensando en estrategias para mejorarlo (cambiar el modelo, cambiar los hiperparámetros, cambiar las variabes elegidas). A su vez, mientras van preparando todo, el/la mentor/a deberá también indagar sobre qué expectativas de resultados tiene cada grupo en función de los modelos que implementarán.

Actividad 3
Una ventana para mirar: *Cloud Computing*

Una vez que analizamos las unidades de procesamiento, podemos asomarnos por la ventana de *Cloud Computing*. Esta parte será 100% expositiva por parte de el/la mentor/a, y se presentaran los conceptos básicos de 'cloud computing' mediante el apoyo de la PPT. A continuación presentar de forma superficial el servicio sagemaker (https://aws.amazon.com/es/sagemaker/faqs/) de AWS, para que los/las estudiantes puedan ver que existe la posibilidad de trabajar con grandes volúmenes de datos y modelos complejos, sin contar con el poder de procesamiento en nuestras computadoras personales.

Encuentro asincrónico
Se le destinará un espacio de seguimiento a los proyectos finales, cada grupo contará sus avances y el/la mentor/a ayudará a marcar el rumbo de los próximos pasos a seguir.

Materiales necesarios:
> PPT *Cross-validation*
> Notebook *Grid-Search*
> Dataset *Titanic*
> PPT *Cloud Computing*

Semana 3

Objetivos

- Conocer la recucción de la dimensionalidad PCA

Actividad 1
Reducción de dimensionalidad: *PCA*

Encuentro sincrónico
"El Análisis de Componentes Principales (PCA) es uno de los algoritmos de Reducción de Dimensionalidad. En esta técnica, se busca generar un nuevo conjunto de variables a partir de variables antiguas, que son la combinación lineal de variables reales. Un nuevo conjunto específico de variables se conoce como los Componentes Principales. Como resultado de la transformación, el primer componente primario tiene la varianza más significativa posible, y cada elemento siguiente tiene la mayor diferencia de potencial respecto a las variables anteriores.

En este encuentro veremos con una PPT el concepto de reducción de la dimensionalidad y el algoritmo PCA. Este es un tema complejo que deberá regularse su complejidad de acuerdo a cada comisión.
Para poder comprenderlo mejor se utilizará una Notebook para ir ejecutando el código y explicarlo. Primero utilizaremos el dataset *Iris* donde se utiliza PCA para poder visualizar el dataset en 2 dimensiones.
Luego se utilizarán Caras de Scikit-Learn de caras para ir mostrando como al disminuir la dimensionalidad se pierde información y hasta cuando sigue identificandose la cara y cuando la pérdida de información ya no permite observar las diferencias,

Actividad 2
Seguimiento de Proyectos

En el espacio de seguimiento de Proyectos, los grupos deberían avanzar en el entrenamiento de distintos modelos e ir mejorando la performance de los mismo. Se incentivará a los/as estudiantes a realizarse preguntas respecto si los modelos son demasiado bueno, tal vez haya que hacer alguna modificación, si no son lo suficientemente buenos, que se podría realizar para mejorarlo, volviendo a pensar en las decisiones tomadas en la limpieza de datos.

Encuentro asincrónico
Se le destinará un espacio de seguimiento a los proyectos finales, cada grupo contará sus avances y el/la mentor/a ayudará a marcar el rumbo de los próximos pasos a seguir.

Materiales necesarios:
> PPT *PCA*
> Notebook *PCA*
> Dataset *Iris*

Semana 4 y Semana 5

Objetivos

- Conocer temáticas de interés de los/as estudiantes
- Repasar temas que hayan quedado dudas
- Terminar Proyectos y armar presentaciones

Actividad
Ventana para mirar a interés de estudiantes

Encuentro sincrónico

Aquí la idea es brindar un espacio de acuerdo a las necesidades de la comisión. Puede ser utilizado para repasar aspectos ya vistos que hayan quedado dudas, para dedicarle el tiempo al trabajo en los Proyectos y/o para finalizar el entrenamiento de modelos y armar la Presentación.

Además, con el concepto "una ventana para mirar" el objetivo de estas clases es poder conocer temáticas que hayan sido planteada de interés por los/as estudiantes. Algunos materiales estarán propuestos pero dependerá del/la mentor/a qué temas hayan surgido a lo largo de las clases o que sirvan a los temas elegidos para los Proyectos.

Algunas opciones son:
- Otros modelos de Ensamble
- Análisis de sentimiento
- Redes neuronales
- Series de tiempo
- Web scraping
- Clasificación de imágenes
- Procesamiento de Lenguaje Natural (NLP)
- Datos georreferenciados
- Pipeline
- Estandarización/Normalización de datos
- Estadística y probabilidad
- Regresión logística

Encuentro asincrónico
Se le destinará un espacio de seguimiento a los proyectos finales, cada grupo deberá terminar sus Notebook y armar una Presentación sobre lo realizado.

Materiales necesarios:
> Material a selección del/la Mentor/a

Semana 6

Objetivos

- Realizar las presentaciones grupales.
- Un resumen de lo aprendido
- Generar un intercambio sobre lo aprendido a lo largo del año

Encuentro sincrónico

Actividad 1
Presentaciones

En este encuentro se realizarán las presentaciones grupales del trabajo realizado para este módulo. La idea de este encuentro es que sea algo festivo, divertido, se podrán utilizar memes y referencias graciosas en las presentaciones para poder sacarle lo solemne, sin perder de vista que es una instancia que es necesario aprobar correctamente.
Se evaluará que las presentaciones tengan sustento en el trabajo realizado durante los dos Módulos, que hayan podido completar y comentar las notebooks para corroborar que entendieron el código de las mismas.

Actividad 2
Repaso de los visto

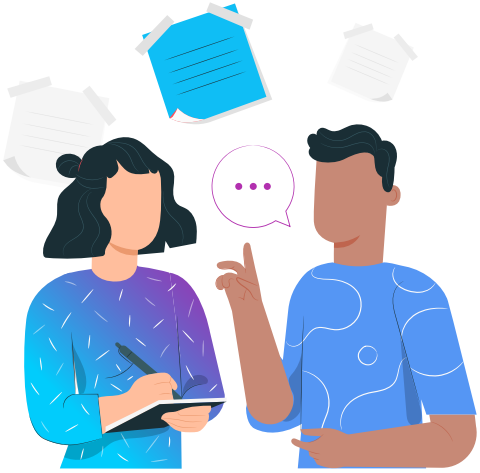
Se realizará una actividad lúdica (como por ejemplo un Kahoot **[2]**) para realizar un repaso de lo visto hasta el momento en el Módulo III para poder reforzar conceptos y ver las dudas que puedan surgir.

Actividad 3
Cierre del curso

Mediante una actividad lúdica se pondrán en juego los diversos aprendizajes del año, buscando que cada estudiante pueda evaluar los contenidos, dinámicas y su propio trayecto y que pueda ponerlo en palabras luego.

La dinámica propuesta es mediante la herramienta de dibujo en la pantalla (o en Miro). El/la mentor/a dirá una frase y cada estudiante deberá poner su nombre de un lado de la línea en la pantalla. De lado izquierdo es "Muy de acuerdo", del lado derecho es "Completamente en desacuerdo."

Se dirán varias frases y cada estudiante deberá posicionar su nombre entre esos dos extremos (pudiendo utilizar el medio o espacios más neutrales también).



"Nunca pensé que podría entender lo que es la Ciencia de Datos"
"Sigo sin entender lo que es la Ciencia de Datos"
"Disfruté mucho de este año de aprendizaje"
"Pude hacerme amigos/as en el proceso"
"Me sentí útil en varias ocasiones"
"Me sentí inútil en varias ocasiones"
"Siempre se puede aprender más"
"Me gustaría seguir profundizando"
"Python me resultó fácil"

Y luego tendrán espacio en Miro o Mural para hacer escritura libre en post-its sobre las siguientes categorías:

- *Cosas que aprendí*
- *Cosas que no terminé de entender*
- *Temas que me gustaría seguir aprendiendo*
- *Me gustó mucho el proyecto de...*
- *Mi clase preferida fue...*
- *Mi clase menos preferida fue...*

Actividad 4
Devoluciones grupales e individuales

Por último se propone un espacio donde el/la mentor/a pueda realizar devoluciones grupales e individuales, siempre en un marco de respeto y sin exponer a nadie frente al grupo.
Finalizando con una reflexión común sobre la experiencia y lo aprendido.

[1] https://kaggle.com/

[2] https://kahoot.com/