Proyecto Integrador.
Programa de Data Science en Digital House.
Entregables + Timeline

Introducción

El Proyecto Integrador (PI) debería representar un aporte original y significativo, aplicando técnicas de data science a un problema interesante. El proyecto integrador se realiza en grupo (excepto que el participante realmente prefiera hacerlo individual).

Identificá un problema relacionado con datos en **tu campo profesional o en una disciplina que realmente conozcas y te apasione**. Si tenés un fuerte interés por el tema de estudio, vas a crear un mejor proyecto y para vos será más satisfactorio realizarlo.

Abajo encontrarás una sección con **Proyectos de Ejemplo** que pueden ayudarte a estimular tu creatividad. Sos bienvenido a usar datos públicos o datos privados, aunque en este último caso tenés que ser cuidadoso con lo que publicás.

Participar en una competencia de **Kaggle** (incluyendo competencias pasadas) también es una opción de proyecto, en cuyo caso los datos te serán provistos por la plataforma.

Comienzo de Proyecto Integrador (Comienza en clase 5)

Discutí tus ideas para el proyecto
Hacia la mitad del curso deberías haber hablado con un miembro del equipo de Data Science acerca de la idea para tu proyecto. Podemos ayudarte a elegir entre diferentes ideas, aconsejarte sobre el alcance apropiado para tu proyecto y asegurar que la pregunta que te formules pueda ser respondida razonable y oportunamente usando las herramientas y técnicas de data science que vemos en el curso.

Para ordenar la charla con nosotros, vas a crear un documento informal, que no es necesario entregar. Debería incluir un breve texto respondiendo a estas preguntas:

¿Cuál es la pregunta que deseás responder?
¿Qué datos estás planeando usar para responder esa pregunta?
¿Qué sabés de los datos hasta ahora?
¿Por qué elegiste este tema?

Características de una buena pregunta para tu proyecto:

Claramente definida: la pregunta puede resumirse fácilmente en una sentencia simple (“¿Podemos predecir A basándonos en B?” o “¿Cuáles son los factores que mejor predicen A?”).

Tan simple como sea posible: la pregunta tiene un foco estrecho en lugar de tener objetivos generales.

Disponibilidad razonable de los datos: la pregunta depende de datos que están disponibles en cantidades significativas y tiempos oportunos.

Hipótesis razonable: la pregunta examina factores (B) que efectivamente podrían ser predictivos de la salida (A).

Entregables

1. Charla Relámpago | Planteo del Problema + Datasets

. Tres a Cinco minutos de presentación que describe tu formulación del problema, define objetivos y criterios de éxito, discute audiencia(s) potencial(es) e identifica al menos 1 o 2 datasets potenciales

. **Formato**: Presentación + Datasets

Link a la consigna detallada

2. Análisis Exploratorio | Diseño y Documentación

. Esta será una descripción de tu enfoque con un resumen bien articulado que incluye:
- . tu objetivo específico,
- . métodos y modelos a utilizar,
- . conjeturas, suposiciones y riesgos,
- . revisiones a tus hipótesis y objetivos iniciales,
- . y un resumen de tu análisis exploratorio.

. Crear una base de datos para tu dataset, si es necesario.

. Describir cualquier manipulación de los datos necesaria y crear un diccionario de datos para tu(s) dataset(s).

. **Formato**: Jupyter Notebook + Presentación en clase con la Notebook

Link a la consigna detallada

Presentación del borrador del proyecto

Darás una breve presentación en el aula acerca del trabajo que hayas realizado hasta el momento, así como sobre tus planes sobre cómo avanzar. Tus slides, código, datos y

visualizaciones deberían estar disponibles en tu repositorio. Aquí tienes algunas preguntas que deberías encarar en tu presentación:

- ¿Qué datos recolectaste y cómo lo hiciste?
- ¿Qué parte de los datos limpiaste y qué partes todavía necesitan limpieza?
- ¿Qué pasos seguiste para explorar los datos?
- ¿Qué primeras conclusiones o percepciones obtuviste de tu exploración?
- ¿Serás capaz de responder a tu pregunta con estos datos? ¿Necesitás recolectar más datos o reformular la pregunta?
- ¿Cómo podrías usar el modelado para responder tu pregunta?

Logística y tips para una entrega exitosa:

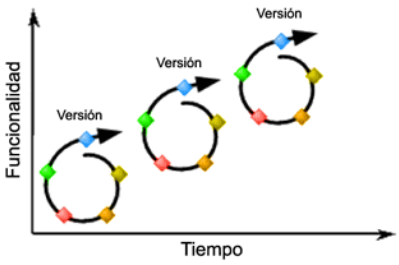
- Por favor, subir un link a tu repositorio (con el material solicitado) a moodle. Realizarás la presentación desde la laptop del aula. No enviar por slack u otro medio a no ser que tengas problema con la plataforma moodle.
- Todos presentarán desde la misma computadora, así que tu presentación debe estar en un formato fácilmente accesible (PDF, PowerPoint, Google Slides, IPython Notebook).
- Tendrás exactamente 5 minutos para presentar, seguido de 1 minuto para preguntas.
- Asegurate de que la pregunta de tu proyecto esté bien clara para todos los participantes en el aula, durante el primer minuto de presentación.
- Contá tu historia de forma interesante y atrapante. Mostrá un ejemplo que ayude a la audiencia a relacionarse y comprometerse con tu tema de estudio.
- Es crítico que practiques la performance de tu presentación y que la cronometres.
- Si ves que tu presentación dura más que 5 minutos, la solución no consiste en hablar más rápido. En cambio, enfocá tu presentación en los aspectos más interesantes de tu proyecto.

Nota al incluir tus datos:

No es práctico que incluyas tu dataset completo en tu repositorio GitHub. Es mejor que incluyas un link a tu fuente de datos y que sólo incluyas una muestra del dataset en tu entrega. (GitHub tiene un límite de 100 MB por archivo y 1 GB por repositorio).

Si tus datos son privados, podés o bien incluir una versión *anonimizada* de tus datos o crear un repositorio GitHub privado (esto último es un servicio pago).

3. Iteración | Análisis + Feedback



. Documentá tu trabajo y obtené feedback sobre tu progreso.

. Creá un **Reporte de Avance** que describa tus resultados iniciales y obstáculos o lecciones aprendidas a lo largo de de estos primeros pasos en el camino.

. Tu reporte debe incluir análisis actualizados tanto visuales como estadísticos de tus datos.

. Te vas a reunir con los instructores para recibir feedback sobre tu trabajo.

. **Formato:** Jupyter Notebook + Discusión en clase

Link a la consigna detallada

Proyecto a entregar:

Una versión avanzada de tu documento de proyecto, junto con los datos, código bien comentado, y visualizaciones. Se debe escribir con una audiencia técnica en mente. Tu trabajo debe incluir los siguientes componentes:

- Declaración del problema e hipótesis
- Descripción de tu conjunto de datos y cómo se obtuvo
- Descripción de los pasos de preprocesamiento que tomaste
- Los nuevos hallazgos realizados al explorar los datos, incluidas las visualizaciones
- Cómo elegiste las features que planéas utilizar en su análisis
- Detalles del proceso de modelado, incluyendo cómo seleccionaste los modelos y cómo fueron validados
- Tus retos y éxitos
- Posibles extensiones o aplicaciones de negocio de su proyecto
- Conclusiones y aprendizajes clave

Tus compañeros y el equipo de instrucción estarán proporcionando feedback. Sin embargo, el informe debe ser autocontenido, y no debe depender de que el lector recuerde tu primera presentación. Cuanto más fácil sea tu artículo, más feedback recibirás! Además, si tus revisores pueden ejecutar tu código en los datos proporcionados, podrán darte una mejor devolución sobre tu implementación.

20/10: Revisión de Pares (individualmente, cada participante)

Vas a proveer feedback en el proyecto de dos de tus colegas, de acuerdo a **peer review guidelines**.

4. Reporte Técnico | Modelo Final + Análisis

. Documentá tus resultados y conclusiones para **colegas y stakeholders técnicos**, incluyendo

. un Sumario Ejecutivo,
. identificación de outliers,
. descripción de cómo definiste las variables,
. discusión de la selección del modelo y su implementación,
. revisión del dataset con visualizaciones y análisis estadístico,
. interpretación de los resultados y su relación con los objetivos iniciales y métricas de éxito del proyecto,
. explicación de pipeline de datos,
. descripción del código fuente utilizado para conducir el análisis,
. recomendaciones a los stakeholders
. y cualquier sugerencia sobre próximos pasos a seguir.

.**Formato:** Jupyter Notebook :: **Entrega Final**

Link a la consigna detallada

5. Presentación al público | Sumario + Recomendaciones

. Crea una **presentación de 10 a 15 minutos** que resalte los insights más importantes de tu proyecto a una audiencia de stakeholders no técnicos.

. Deberías mencionar tus objetivos, métodos, conclusiones y recomendaciones, representadas con visualizaciones, gráficos y técnicas de storytelling.

. Incluí un apéndice que resuma los aspectos más técnicos de tu trabajo.

. Formato: Presentación / PDF

Link a la consigna detallada

Presentación final del proyecto y el informe

El repositorio de tu proyecto en GitHub debería contener lo siguiente:

- Documento de proyecto: cualquier formato (PDF, Markdown, etc.)
- Slides de presentación: cualquier formato excepto Keynote (PDF, PowerPoint, Google Slides, IPython Notebook, etc.)
- Código: scripts de Python comentados, y cualquier otro código que utilizaste en el proyecto
- Visualizaciones: integradas en tu documento y/o diapositivas
- Datos: archivos de datos en formato "raw" o "procesado"
- Diccionario de datos (aka "code book"): descripción de cada variable, incluidas sus unidades

Logística y consejos para el éxito (similar a la última vez)

- Envíe un link a su repositorio (con diapositivas) antes de las 6 de la tarde del día que se presente.
- Independientemente del día que esté presentando, su repositorio también debe contener los otros componentes requeridos del proyecto antes de las 6 pm el último día de clase.
- Tendrás exactamente 12 minutos para presentar, seguido de 2 minutos de preguntas. Practique su presentación y tiempo usted mismo!
- Su presentación debe comenzar con una recapitulación de la información clave de la presentación anterior (incluida la pregunta del proyecto), pero debe dedicar la mayor parte de su presentación a discutir lo que ha ocurrido desde entonces.
- Si su presentación es demasiado larga, enfoque en torno a los aspectos más interesantes de su proyecto, en lugar de tratar de incluir todos los detalles.
- Cuente su historia de una manera atractiva.
- Te proponemos invitar a sus amigos y familiares a asistir.

Recursos útiles para empezar

Proyectos de Ejemplo

https://github.com/ahlusar1989/DAT-project-examples
https://gallery.generalassemb.ly/DS
https://gallery.generalassemb.ly/DSI?metro=

Fuentes Públicas de Datos

https://github.com/justmarkham/DAT8/blob/master/project/public_data.md
Quandle
INDEC
https://datos.gob.ar/
https://datar.noip.ar

Competencias de Ciencias de Datos

Kaggle, DrivenData, CrowdANALYTIX, TunedIT, InnoCentive