IMT 2200 - Introducción a Ciencia de Datos

Pontificia Universidad Católica de Chile
Instituto de Ingeniería Matemática y Computacional
Semestre 2023-1

Profesor: Rodrigo A. Carrasco

Proyecto de Ciencia de Datos

Este documento detalla el proyecto semestral. Las fechas límite importantes son:

- Entrega de Propuesta: jueves 12 de octubre a las 23:59 hrs.
- Entrega del repositorio: jueves 23 de noviembre a las 23:59 hrs.
- Entrega de la página web: martes 28 de noviembre a las 9:00 hrs.
- Archivo con la presentacion: martes 28 de noviembre a las 9:00 hrs.
- Presentaciones: martes 28 y jueves 30 de noviembre.

1. Objetivo

El objetivo de este proyecto es integrar los conceptos estudiados a lo largo de todo el curso. Para ello deben desarrollar el proceso completo de ciencia de datos para resolver una pregunta sobre algún tema a su elección. Esto comprende la adquisición y procesamiento de datos, análisis exploratorio, visualización de datos, análisis estadístico, modelamiento y comunicación de resultados.

El problema a abordar puede estar relacionado a cualquier tema de interés científico, social o comercial que ustedes tengan. Para que su proyecto pueda ser desarrollado exitosamente, es importante que la pregunta a resolver sea suficientemente específica, cuantificable (de manera que pueda ser abordada en base a datos), accionable (es decir, que la respuesta sirva para definir acciones o decisiones) y relevante. Además, el problema debe ser factible de resolver dentro de un tiempo acotado. Puede encontrar un buen resumen de estos puntos en el siguiente artículo:

https://towardsdatascience.com/7-steps-to-a-successful-data-science-project-b452a9b57149

2. Metodología de trabajo.

- El proyecto se desarrollará en los grupos de trabajo definidos al inicio del semestre. Las entregas consistirán de un documento de propuesta, una presentación final, y una entrega del desarrollo completo en un sitio web armado por ustedes.
- El proyecto debe ser desarrollado en lenguaje Python, y utilizar datos de acceso públicos o privados, siempre que se respeten las políticas de uso y publicación de los mismos.
- Para el desarrollo y entregas del proyecto se utilizará la plataforma GitHub. Cada equipo deberá crear un repositorio público que albergará los datos y códigos del proyecto, y una página web para comunicar sus resultados, creada mediante GitHub Pages (https://pages.github.com).
- Este repositorio será el **único canal oficial de entrega del proyecto**, y las correcciones se realizarán sobre la última versión modificada antes del plazo definido para cada entregable.
- Durante el semestre, se ha habilitado en Canvas un foro de discusión específico para el proyecto, para canalizar todas las dudas que surjan en el curso. Además, en caso de requerirlo se podrán agendar horas de consulta con el profesor o las ayudantes del curso.

3. Entregables.

3.1 Propuesta de proyecto (jueves 12 de octubre a las 23:59 hrs.)

Cada equipo podrá definir su propio tema y datos de trabajo en torno a los cuales deberá plantear un problema a resolver en base a ciencia de datos. Como primer entregable, cada equipo deberá generar una **Propuesta del proyecto** a desarrollar, la cual consta de una propuesta escrita, y una muestra de los datos a utilizar. La propuesta escrita debe incluir los siguientes puntos:

- **Título del proyecto.**
- **Contexto y motivación**: discuta la motivación y razones para su elección del proyecto, incluyendo el contexto y alguna investigación que haya incluido en su decisión. Comente también sobre la audiencia objetivo de su análisis (quién será el tomador de decisiones que se apoyará en su trabajo).
- **Objetivos**: describa los objetivos científicos y de inferencia de su proyecto. ¿Qué pregunta espera resolver o responder, y para qué? ¿Cuáles son los beneficios o accionables que podría lograr mediante este proyecto? ¿Cuál es la audiencia objetivo de su análisis?
- **Datos**: describa los datos a utilizar, sus características (variables, tipo, formato, volumen), origen y forma de recolección. Los datos pueden provenir de fuentes privadas (siempre que no tengan restricciones de uso o publicación), de fuentes públicas (gobierno, plataformas de datos abiertos, web, etc.) o pueden ser recolectados mediante algún método propuesto por su equipo (Ej. web scrapping, Twitter, RRSS, etc.).
- **Preguntas de investigación**: proponga al menos 5 preguntas específicas que buscarán responder con base en los datos presentados. Estas preguntas pueden ser relativas a relaciones entre variables, capacidad de predicción de una variable, clasificación, identificación de clusters o categorías, etc.
- **Diseño tentativo**: explique brevemente los métodos computacionales y estadísticos que espera usar en su análisis.

La parte escrita puede ser entregada en formato PDF como un informe. Además, deben incluirse los archivos de datos crudos con los que se propone trabajar, o al menos una muestra de ellos. Todo ello debe quedar disponible en el repositorio del proyecto.

El objetivo de esta propuesta es validar y eventualmente ajustar la propuesta de trabajo de cada equipo, que puede ser modificada a lo largo de su desarrollo.

3.2 Repositorio GitHub, Jupyter Notebook y Códigos (jueves 23 de noviembre a las 23:59 hrs.)

Cada grupo deberá entregar un compilado de los datos y códigos utilizados en su desarrollo, de manera que éste sea completamente replicable, desde los datos crudos hasta el producto final. Los códigos deben estar escritos en forma ordenada y eficiente, con comentarios para enteder qué hace cada parte, de manera que sean fácilmente legibles y reproducibles. Si no se entiende lo que hace una parte del código puede ser evaluado negativamente.

El desarrollo debe estar estructurado en forma de un Jupyter Notebook, que incluya los siguientes elementos:

- **Contexto y motivación:** describa el contexto y objetivos del proyecto, para una audiencia que no tiene información previa respecto a él. Incluya referencias, artículos o discusiones que le hayan servido de motivación.
- **Preguntas objetivo:** ¿cuáles son las preguntas que busca responder mediante su análisis? ¿Cómo han evolucionado sus preguntas a lo largo del proyecto, y si han aparecido otras nuevas en el camino?
- **Datos:** describa los datos utilizados y su origen, y documente los procedimientos de recolección, preparación y transformación de datos. Describa los datos indicando el tipo de datos, su estructura y los elementos que se describen en la bae de datos.
- **Análisis exploratorio de datos:** describa el proceso de exploración de sus datos, incluyendo visualizaciones, análisis gráficos, análisis estadísticos, etc. Explique los resultados y conclusiones preliminares obtenidas, y cómo estos influyen o motivan el método de modelamiento elegido. Justifique todas las decisiones adoptadas, tanto la elección de herramientas como las elecciones de visualización.
- **Análisis de datos:** describa y documente los métodos computacionales y estadísticos aplicados al modelamiento de sus datos. Justifique las elecciones y decisiones adoptadas (ej: selección de variables y parámetros, configuración de algoritmos, normalizaciones, validaciones, etc.), y evalúe sus resultados en base a métricas apropiadas al modelo elaborado (Ej. error cuadrático, intervalo de confiabilidad, precisión, recall, F1, matriz de confusión, etc.)
- **Resumen de los resultados:** explique narrativa y visualmente sus resultados, cómo estos responden a la pregunta original.
- **¿Qué podría salir mal?:** explique las posibles limitaciones o sesgos de sus datos o análisis. Identifique qué problemas se podrían generar a partir de su solución o de las decisiones que esta recomiende.

El Notebook, datos y códigos asociados deberán estar disponible en el repositorio del proyecto al acumplirse el plazo de entrega estipulado. Todo código o documento subido posteriormente a la fecha límite no será considerado en la evaluación de la entrega.

3.3 Página web del proyecto (martes 28 de noviembre a las 9:00 hrs.)

Asociado al mismo repositorio, cada equipo deberá crear una página tipo GitHub Pages para comunicar y publicar los resultados de su proyecto. Esta página tiene como objetivo relatar en forma efectiva los principales objetivos y hallazgos del proyecto, incluyendo texto narrativo y figuras, en un lenguaje y formato adecuado para público general. Los códigos y datos estarán enlazados a la página web.

3.4 Presentación (martes 28 y jueves 30 de noviembre)

- Cada equipo realizará una presentación de **5 minutos** de duración que tendrá otros **5 minutos** de pregutnas. Esta presentación debe enfocarse en las contribuciones y resultados del proyecto más que en los aspectos técnicos, y destacar los hallazgos y conclusiones del proyecto, o mensajes centrales que la audiencia debería recordar.
- Cada equipo puede escoger quién realiza la presentación, pero todos deben respondaer preguntas respecto al proyecto.
- La asistencia las presentaciones es **obligatoria** para todos los estudiantes y será considerada como de las Actividades del Curso.
- El documento a presentar debe ser subido a Canvas a más tardar el martes 28 de noviembre a las 9:00 hrs.

4. Requisitos generales.

- **Datos:** los datos del proyecto pueden obtenerse de cualquier fuente pública o privada, pero su extracción debe estar debidamente documentada en el código. Su proyecto debe considerar al menos **2 datasets** distintos, que se puedan combinar para realizar el análisis propuesto. El dataset combinado debe tener *al menos* 7 variables relevantes para el análisis y al menos 100 registros. **No** está permitido usar datos de repositorios de proyectos de Ciencia de Datos como: Kaggle, UCI Machine Learning Repository, Datacamp, etc., pero puede ir a las fuentes de algunos de esos sitios y obtener los datos de esa fuente original.
- **Análisis de datos:** En la sección de *Análisis de Datos* debe incluirse al menos uno de los modelos o algoritmos que veremos en el curso: regresión, clasificación, clustering. El análisis de resultados debe considerar las métricas típicamente utilizadas para cada tipo de algoritmo. No se evaluará la precisión de los modelos desarrollados, sino la coherencia, rigurosidad, buena implementación y evaluación de los mismos.
- **Cuestionamiento:** En la sección de *¿Qué podría salir mal?* asociados al proyecto, deben cuestionar los objetivos y resultados de su proyecto desde un punto de vista ético y asociado a los problemas que podría generar su solución o recomendación. Pueden usar ejemplos asociados al libro *Weapons of Math Descruction* para aterrizar algunos problemas que puedan aparecer.
- **Códigos**: todos los códigos deben estar en lenguaje Python, escritos en forma ordenada y siguiendo buenas prácticas de programación (eficiencia, reusabilidad, modularización, uso de anotaciones, etc.). Se evaluará la calidad y estructura de los Notebooks o códigos asociados. Todo el código entregado debe ser escrito únicamente por los integrantes del grupo. Cualquier elemento no escrito por ustedes, debe ser indicado cláramente y no será considerado en la parte de "Novedad y Trabajo" de la rúbrica de evaluación.

5. Evaluación.

La nota final del proyecto se calculará en base a la ponderación de los cuatro entregables:

- **Propuesta del proyecto:** 20%.
- **Jupyter Notebook y Códigos:** 50%.
- **Página web:** 10%.
- **Presentación final:** 20%.

Para cada entregable, se publicará una rúbrica detallada con los elementos a incluir y evaluar.