Pontificia Universidad Católica de Chile
Instituto de Ingeniería Matemática y Computacional
Semestre 2025-2

IMT2200 - Introducción a Ciencia de Datos

Proyecto de Ciencia de Datos

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

- Entrega de Propuesta: 11 de septiembre a las 23:59 hrs.
- Entrega Inicial del Repositorio: 21 de octubre a las 23:59 hrs.
- Entrega Final Repositorio: 20 de noviembre a las 23:59 hrs.
- Entrega de la Página web y el Video: 27 de noviembre a las 23:59 hrs.
- Entrega Evaluación Videos: 5 de diciembre a las 23:59 hrs.

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 desarrollo del proyecto será realizado en equipos y está diseñado para simular el ciclo completo de un proyecto de ciencia de datos aplicado. Se espera que los estudiantes apliquen de manera integrada los conocimientos adquiridos durante el curso, desde la formulación de una pregunta relevante hasta la entrega de productos reproducibles y comunicables. El trabajo será iterativo y evaluado en distintas etapas, incluyendo entregables técnicos y de comunicación. A lo largo del semestre, los equipos contarán con apoyo docente para orientar sus decisiones metodológicas y técnicas, y deberán mantener una documentación clara y actualizada del proceso seguido.

GitHub ha sido seleccionado como la plataforma oficial de desarrollo y entrega del proyecto debido a su capacidad para gestionar versiones, organizar el trabajo colaborativo y facilitar la reproducibilidad de los análisis. A través del uso de repositorios públicos, los equipos podrán documentar de forma estructurada su progreso, integrar código, datos y visualizaciones, y desarrollar una página web que comunique los resultados de manera accesible. Además, el uso de GitHub Pages permite presentar el producto final como un recurso interactivo y profesional, replicando buenas prácticas de publicación utilizadas en ciencia de datos aplicada.

- 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

El proyecto está estructurado en una serie de entregables progresivos, diseñados para guiar a los equipos en el desarrollo iterativo y reflexivo de su trabajo. Cada entrega cumple un rol específico dentro del proceso de ciencia de datos: desde la formulación clara del problema y la exploración inicial de datos, hasta el análisis, la comunicación de resultados y la evaluación crítica. Esta estructura no solo permite avanzar de manera organizada y recibir retroalimentación oportuna, sino que también refleja cómo se trabaja en proyectos reales, donde el aprendizaje ocurre a través de ciclos sucesivos de exploración, análisis y revisión. El cumplimiento riguroso y documentado de cada etapa será clave para la calidad final del proyecto.

3.1 Propuesta de proyecto (11 de septiembre 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 de 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 debe ser entregada en formato PDF como un informe. Además, debe incluirse una muestra de 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 y en Canvas deben subir (en el módulo correspondiente) el enlace a su repositorio que debe estar en formato público.
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 Entrega Inicial del Repositorio (21 de octubre a las 23:59 hrs.)

Esta entrega tiene como objetivo consolidar el trabajo realizado hasta la etapa de análisis exploratorio de datos (EDA), dejando documentado el contexto del proyecto, las preguntas que guían el análisis, las fuentes de datos utilizadas y una primera aproximación a su exploración. Esta entrega es fundamental, ya que permite validar tempranamente la dirección del trabajo, identificar desafíos metodológicos y establecer una base sólida sobre la cual desarrollar los modelos y análisis posteriores.
Todo el material debe quedar registrado en el mismo repositorio de GitHub del equipo, incluyendo el notebook con una narrativa clara y bien estructurada del proceso hasta este punto.
El desarrollo debe estar estructurado en forma de un Jupyter Notebook, que incluya al menos 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.

El Notebook, datos y códigos asociados deberán estar disponible en el repositorio del proyecto al cumplirse 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 Actualización Final del Repositorio – GitHub, Jupyter Notebook y Códigos (20 de noviembre a las 23:59 hrs.)

En esta entrega, los equipos deberán completar y actualizar su repositorio con el desarrollo completo del proyecto, incorporando los métodos de modelamiento, la evaluación de resultados y las conclusiones obtenidas. El análisis debe continuar sobre la base del trabajo entregado previamente, refinando las preguntas de investigación cuando sea necesario y documentando las decisiones metodológicas tomadas en esta segunda etapa. Se espera un trabajo bien estructurado y reproducible, en el que los datos, códigos y visualizaciones estén organizados de manera clara y coherente.
El desarrollo debe estar estructurado en forma de un Jupyter Notebook, que incluya al menos los siguientes elementos:

- **Resumen del proyecto**: breve introducción que conecte con lo entregado previamente.
- **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 cumplirse el plazo de entrega estipulado. Todo código o documento subido posteriormente a la fecha límite no será considerado en la evaluación de la entrega.

3.4 Página web del proyecto (27 de noviembre a las 23:59 hrs.)

Como parte del desarrollo del proyecto, cada equipo deberá crear una página web utilizando **GitHub Pages**, asociada al mismo repositorio donde se aloja el código. Esta página tiene como objetivo *comunicar los resultados del proyecto de forma clara, atractiva y accesible para una audiencia general*, destacando los hallazgos más relevantes mediante texto narrativo y visualizaciones.

Además de ser un producto comunicacional dentro del curso, esta página puede convertirse en un **valioso recurso para su portafolio personal como futuros científicos o analistas de datos**. Les permitirá demostrar su capacidad para enfrentar un problema real, estructurar un análisis, comunicar resultados y presentar visualmente sus conclusiones. Por ello, se espera una presentación pulida, con un diseño sencillo pero profesional, y un mensaje claro sobre qué se hizo, por qué es importante y qué se aprendió.

La página debe incluir texto, imágenes y enlaces relevantes, pero no necesita reproducir todos los detalles técnicos del proyecto (que ya están en el notebook). Puede ser construida con herramientas simples como `Markdown`, HTML básico o generadores estáticos disponibles en GitHub Pages. Los códigos y datos deben estar correctamente enlazados desde esta página.

3.5 Video con Presentación (27 de noviembre a las 23:59 hrs.)

Además del análisis técnico y la presentación escrita del proyecto, es fundamental que los estudiantes puedan comunicar de manera oral y visual los resultados obtenidos. La capacidad de presentar un proyecto de ciencia de datos de forma clara, estructurada y convincente es una habilidad clave en contextos profesionales, académicos y de divulgación. Por eso, esta entrega consiste en la grabación de un video donde cada equipo resuma y comunique los principales hallazgos de su proyecto, destacando las motivaciones, decisiones y conclusiones más relevantes.

Esta presentación debe estar dirigida a una audiencia general y centrarse en el valor del análisis realizado, más que en los detalles técnicos del código.

- Cada equipo entregará un video con una presentación de 10 minutos de duración máximo. 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.
- Todos en el equipo deberán presentar y se entregará una nota individual por la presentación.
- Cada alumno deberá evaluar tres videos posteriormente, con una rúbrica que entregaremos.
- El enlace al video debe ser subido a Canvas a más tardar el 27 de noviembre a las 23:59 hrs.

3.6 Trabajo Individual de Análisis Crítico (5 de diciembre a las 23:59 hrs.)

Como parte de la evaluación individual del proyecto, cada estudiante deberá revisar y evaluar tres videos presentados por otros grupos del curso. Esta actividad tiene como objetivo fomentar una reflexión crítica sobre la comunicación de resultados en ciencia de datos, así como promover el aprendizaje a través de la observación de diferentes enfoques, soluciones y estilos de presentación.

Cada evaluación deberá completarse utilizando la **misma rúbrica oficial que será utilizada por el equipo docente para evaluar los videos de presentación de cada grupo**. Esta rúbrica incluirá aspectos como claridad en la comunicación, profundidad de los resultados, relevancia de las conclusiones y calidad visual del material. Además de asignar puntajes, cada estudiante deberá entregar una retroalimentación escrita breve y constructiva para cada video revisado.

Los tres videos que deberá evaluar cada estudiante serán **asignados aleatoriamente por el equipo docente**, con el objetivo de asegurar diversidad de temas y una distribución equilibrada de las evaluaciones. La calidad y profundidad de estas evaluaciones será considerada como parte de la nota individual del proyecto. El objetivo no es sólo entregar una calificación justa, sino también demostrar capacidad de análisis crítico y de comunicación en el contexto del trabajo de otros equipos.

4 Requisitos generales

Para asegurar la calidad y coherencia del proyecto, existen ciertos requisitos mínimos que deben cumplirse en cuanto a los datos utilizados, el alcance del análisis, la implementación técnica y la reflexión crítica. Estos requisitos tienen como objetivo asegurar que cada equipo enfrente un desafío realista pero no trivial, que les permita aplicar los conceptos del curso en su totalidad. También buscan fomentar buenas prácticas de programación, documentación y comunicación, que son esenciales en cualquier proyecto de ciencia de datos.

A continuación, se detallan los criterios obligatorios que deben considerarse en el desarrollo del proyecto.

- **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 grupal del proyecto se calculará en base a la ponderación de los cuatro entregables:

- Propuesta: 20%
- Repositorio con el análisis del problema: 20%
- Actualización del Repositorio con Resolución: 40%

- Página web y Video de Presentación: 20%

Para cada entregable, se publicará una rúbrica detallada con los elementos a incluir y evaluar.
La parte individual del trabajo tiene como objetivo evaluar la comprensión personal del proceso completo del proyecto. Deberá ser entregado por cada estudiante de manera independiente y reflejará su análisis, contribución y aprendizaje durante el desarrollo del proyecto grupal.
Para esta parte individual, ustedes deberá evaluar 3 videos con una rúbrica que entregaremos, haciendo una crítica constructiva del trabajo. La crítica será evaluada y su nota afectará en forma individual la nota del proyecto como se indica en el programa del curso.

6 Integridad Académica

La realización del proyecto debe regirse por los más altos estándares de integridad académica. Todo el trabajo entregado, incluyendo código, visualizaciones, análisis y reflexiones escritas, debe ser desarrollado exclusivamente por los integrantes del equipo. No se permite reutilizar trabajos de otros cursos, copiar o adaptar código ajeno sin citar, ni presentar como propias respuestas generadas por herramientas automatizadas sin indicarlo de forma explícita.

El uso de herramientas de inteligencia artificial generativa (como ChatGPT, Copilot u otras) está permitido exclusivamente como apoyo para el desarrollo de ideas o fragmentos de código. En caso de utilizar IA generativa, se debe incluir la cita correspondiente indicando el *prompt* utilizado. Todo el contenido escrito que no sea código debe ser de autoría exclusiva del equipo, redactado con comprensión y reflexión propia.

El uso de estas herramientas está estrictamente prohibido durante interrogaciones y actividades en clase. El incumplimiento de estas normas será considerado una falta grave a la integridad académica y será sancionado conforme a los reglamentos de la Universidad y del Instituto de Ingeniería Matemática y Computacional.

6.1 Buenas prácticas y ejemplos

A continuación entregamos algunos ejemplos para que se guíen en el uso debido de tecnología y fuentes. Si tienen dudas, deben preguntar a los ayudantes o el profesor del ramo.

Práctica aceptada	**Práctica prohibida**
Usar ChatGPT para obtener una idea general de cómo implementar un algoritmo, e incluir al final del notebook: "Parte del código fue inspirado por una respuesta generada en ChatGPT con el prompt: *"How to implement KMeans in Python with sklearn?"*"	Copiar directamente código completo desde ChatGPT o Copilot sin entenderlo ni citar la fuente.
Redactar personalmente las conclusiones y reflexiones del proyecto, basándose en los resultados obtenidos por el equipo.	Usar ChatGPT para redactar la sección de conclusiones del notebook o del informe individual y presentarla sin cambios ni revisión.
Consultar documentación oficial de bibliotecas como pandas o scikit-learn y adaptar ejemplos con comprensión.	Copiar y pegar código de Stack Overflow o tutoriales sin modificaciones ni explicación del funcionamiento.
Citar fuentes externas (páginas web, artículos, notebooks públicos) cuando se adaptan ideas o estructuras del código.	Utilizar código externo como si fuera propio, omitiendo la fuente o presentándolo como desarrollo original.