Pontificia Universidad Católica de Chile
Instituto de Ingeniería Matemática y Computacional
Semestre 2021-2

IMT2200 - Introducción a Ciencia de Datos

Programa del Curso

1 Información General

- **Profesora**: Paula Aguirre Aparicio (paaguirr@ing.puc.cl)
- **Ayudantes**:
  - Ayudante de Cátedra: Vicente Agüero (vicenteaguero@uc.cl)
  - Ayudante Corrector: TBD
- **Horarios**:
  - Cátedra: M-J:2
  - Ayudantía: W:3
  - Horario de consultas profesora (por Zoom o presencial): W 16:00 - 17:00 hrs.
- **Página web**: `https://imt2200.github.io/`

2 Descripción del Curso

Las organizaciones utilizan sus datos para apoyar la toma de decisiones, y para desarrollar productos y servicios intensivos en datos. El conjunto de competencias requeridas para apoyar estas funciones se ha agrupado bajo el término Ciencia de Datos. En este curso los estudiantes analizarán la importancia de este campo y su crecimiento exponencial, describiendo sus principios básicos y las principales técnicas y herramientas utilizadas. Los estudiantes aprenderán sobre recolección e integración de datos, análisis exploratorio de datos, análisis descriptivo y predictivo, y creación de productos de información.

3 Resultados de Aprendizaje

Al finalizar este curso, los estudiantes habrán logrado los siguientes aprendizajes:

1. Describir lo que es la ciencia de datos, y su importancia para la ciencia, sociedad y negocios.
2. Identificar los conjuntos de habilidades necesarios para ser un científico de datos.
3. Identificar problemas éticos y de privacidad que emergen en ciencia de datos.
4. Explicar las etapas y tareas que forman parte del ciclo de vida de un proyecto de ciencia de datos.
5. Reconocer distintos tipos y formatos de datos estructurados y no estructurados.
6. Desarrollar el proceso de extracción, transformación y carga de datos para un proyecto sencillo de ciencia de datos.
7. Identificar los fundamentos matemáticos y estadísticos para el análisis exploratorio y modelamiento estadístico en un proyecto de ciencia de datos.

8. Utilizar técnicas de análisis exploratorio y modelamiento apropiados para problemas sencillos de ciencia de datos.
9. Aplicar algoritmos básicos de aprendizaje de máquina para el análisis descriptivo y predictivo en proyectos sencillos de ciencia de datos.
10. Comunicar efectivamente los resultados de un proyecto de ciencia de datos.

4 Contenidos

A lo largo del semestre, estudiaremos los siguientes contenidos:

1. Introducción
2. Extracción, transformación y carga de datos.
3. Fundamentos matemáticos y estadísticos para el análisis de datos.
4. Análisis Exploratorio de Datos.
5. Modelamiento estadístico.
6. Introducción a algoritmos de aprendizaje de máquinas.
7. Visualización y comunicación de datos y resultados.

5 Metodología del Curso

5.1 Clases y ayudantías

Este curso consta de dos cátedras y una ayudantía semanal. Si bien no existen requisitos formales de asistencia, la participación en cada una de estas instancias es crucial para aprender el material presentado en este curso, y se espera que los estudiantes asistan a todas ellas en lo posible.

Las clases comenzarán en formato remoto y pasarán a ser **presenciales a partir del 20 de septiembre de 2021**, si las condiciones sanitarias lo permiten, y siguiendo todas las directrices y protocolos establecidos por la Universidad y publicados en `https://www.uc.cl/uc-contra-el-coronavirus/` . Las ayudantías se mantendrán en formato remoto durante todo el semestre.

Las clases tendrán en general una parte expositiva, y otra de ejercicios de programación en vivo. Todo el material requerido para la clase estará disponible antes del comienzo de ésta en la página web y repositorio GitHub del curso. En el caso de clases presenciales, se recomienda idealmente traer un computador personal para seguir el desarrollo de ejercicios de programación. En todo el curso, se utilizará en lenguaje de programación Python.

5.2 Materiales del curso

Todo el material actualizado del curso estará disponible para los estudiantes en la página y repositorio GitHub del curso. Es responsabilidad de cada estudiante revisar frecuentemente la página del curso, y descargar oportunamente el material (slides, notebooks, datos) requeridos para cada actividad.

5.3 Anuncios y comunicaciones

Todos los anuncios oficiales del curso de realizarán a través de la plataforma Canvas y se publicarán en la página web. Es responsabilidad de cada estudiante revisar frecuentemente estos medios y mantenerse actualizado de las novedades del curso.
Para situaciones puntuales o dudas de cualquier tipo, por favor comunicarse con la profesora y/o ayudantes del curso vía mail.

5.4 Pandemia y presencialidad

La planificación del paso a clases presenciales a partir del 20 de septiembre de 2021 responde a la situación sanitaria favorable que vivimos a comienzos del semestre 2021-1. Sin embargo, el último año nos ha demostrado que es difícil predecir con exactitud la evolución de la pandemia de COVID-19, y que la realidad actual requiere flexibilidad y capacidad de adaptación en todos los aspectos de nuestras vidas. En caso de que la evolución de la pandemia obligue a realizar modificaciones a la planificación de clases o al programa del curso en general, se informará a los estudiantes con la mayor anticipación posible, velando por que todos puedan participar adecuadamente en todas las actividades de aprendizaje del curso.

Además, cada estudiante, ayudante y profesor puede enfrentar situaciones personales, familiares o de salud particulares, que requieren ser abordadas con empatía y solidaridad. En caso de que algún estudiante enfrente dificultades de cualquier tipo para seguir adecuadamente el curso, por favor comunicarse lo antes posible con el equipo docente para buscar soluciones en conjunto.

6 Evaluaciones

Las evaluación del curso se realizará a través de tres instancias: tareas y controles individuales, y un proyecto grupal.

6.1 Tareas (NT)

Durante el semestre se desarrollarán 4 tareas, consistentes en ejercicios aplicados de programación y análisis de datos. Las tareas son de desarrollo **individual**, y se promedian para dar origen a la nota NT, calculada como:

$$NT = \frac{T_1 + T_2 + T_3 + T_4}{4} \tag{1}$$

donde T_i es la nota obtenida en cada una de las 4 tareas.
Las tareas de programación se entregan únicamente a través de un repositorio privado y personal de cada estudiante. El repositorio se alojará en la plataforma GitHub, y será entregado por el equipo docente. Este el medio de entrega oficial y exclusivo de las Tareas y Proyecto del Curso, no se aceptarán entregas mediante mail o Canvas.

6.2 Controles (NC)

Semanalmente se realizarán controles breves consistentes en preguntas relativas a los contenidos vistos en las clases anteriores. Los controles se realizarán simultáneamente en los últimos 10 minutos de la clase, a través de Canvas o en formato presencial dependiendo de la modalidad de asistencia de cada estudiante. En caso de atraso o inasistencia, **no se puede recuperar el control**, y éste se califica con puntaje 0. Situaciones particulares de inasistencias extendidas por licencia médica u otros se analizarán en forma individual.

Cada control constará de alrededor de 4 preguntas, y se espera realizar en total 12 controles. Cada pregunta tiene puntaje 0 o 1, y los puntajes P_i de todas las preguntas del semestre se acumularán para dar origen a una única nota NC, calculada como:

$$NC = \frac{\sum P_i}{NP} \cdot 6 + 1 \tag{2}$$

Para este cálculo, se considerará solo el **75% superior** de las preguntas contenidas en todos los controles realizados a lo largo del semestre. Por ejemplo, si se realizan 12 controles equivalentes a un total de 48 preguntas, para el cálculo de la nota se eliminan los 12 peores puntajes P_i (25%), *incluyendo inasistencias.*

6.3 Proyecto (NP)

A lo largo del semestre, se desarrollará un proyecto grupal de aplicación, en el cual los estudiantes deberán desarrollar el proceso completo de ciencia de datos para resolver una pregunta sobre algún tema a su elección. Esto comprende el planteamiento del problema, la adquisición de datos, análisis exploratorio, visualización

de datos, análisis estadístico, modelamiento y comunicación de resultados.

El proyecto se compone de tres entregables, cada uno de los cuales será evaluado con una nota y ponderado de acuerdo a los siguientes porcentajes para el cálculo de la nota de proyecto NP:

- Propuesta: 20%
- Repositorio y página GitHub: 60%
- Presentación visual: 20%

Las instrucciones detalladas para la realización del proyecto, forma y fecha de entrega de cada uno de los entregables se explicarán en el enunciado del mismo, que será presentado en clases y publicado en la página del curso.

6.4 Nota final

La nota final del curso se calculará de acuerdo a la siguiente fórmula:

$$NF = 0.5 \cdot NP + 0.3 \cdot NT + 0.2 \cdot NC \tag{3}$$

- Las fechas de tareas y controles estarán detalladas en el Calendario publicado en la página web del curso.
- Todas las notas serán calculadas con dos decimales, salvo la nota final del curso que se calculará con un decimal.
- Cualquier situación de copia detectada en alguna evaluación tendrá como sanción un 1,1 final en el curso. Esto sin perjuicio de sanciones posteriores que estén de acuerdo a la Política de Integridad Académica de la Universidad, que sean aplicables al caso.
- Todas las notas del curso serán publicadas a través de la plataforma Canvas.

7 Bibliografía

7.1 Mínima:

- Cathy O'Neil and Rachel Schutt. *"Doing Data Science, Straight Talk from the Frontline"*. O'Reilly Media, 2013.
- Wes McKinney. *"Python for data analysis: Data wrangling with Pandas, NumPy, and IPython"*. O'Reilly Media, 2da edición, 2017.

7.2 Complementaria:

- Joel Grus. *"Data Science from Scratch: First Principles with Python"*. O'Reilly, 2ª edición, 2019.
- Foster Provost and Tom Fawcett. *"Data Science for Business: What you need to know about data mining and data-analytic thinking"*. O'Reilly Media, 2013.

Además de los libros sugeridos como bibliografía, este curso hace uso de distintos recursos y documentación online, que serán referenciados cuando corresponda en clases y en la página del curso.

8 Integridad Académica

Se espera los alumnos del Instituto de Ingeniería Matemática y Computacional (IMC) de la Pontificia Universidad Católica de Chile mantengan altos estándares de honestidad académica, acorde al **Código de Honor** de la Universidad. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a la iniciación de un procedimiento y a la aplicación de las sanciones contenidas en el Reglamento sobre la Responsabilidad Académica y Disciplinaria de los miembros de la Comunidad Universitaria. Es responsabilidad de cada estudiante conocer estos documentos, disponibles en `https://admisionyregistros.uc.cl/alumnos/novatos/reglamentos-estudiantiles`.

Específicamente, para este curso rige obligatoriamente la siguiente política de integridad académica. Todo trabajo presentado por un alumno para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno, sin apoyo en material de terceros. Por "trabajo" se entiende en general los controles, interrogaciones, las tareas de programación entre otros. En el caso de proyectos grupales, el trabajo presentado por el grupo debe ser elaborado exclusivamente por los integrantes del mismo, sin participación de terceros.

En particular, si un alumno copia un trabajo, o si a un alumno se le prueba que compró o intentó comprar un trabajo individual o grupal, obtendrá nota final 1.1 en el curso y no se le permitirá retirar el curso de la carga académica semestral.

Por "copia" se entiende incluir en el trabajo presentado como propio, partes hechas por otra persona. En caso que corresponda a "copia" a otros alumnos, la sanción anterior se aplicará a todos los involucrados. En todos los casos, se informará al comité académico del IMC para que tome sanciones adicionales si lo estima conveniente. Obviamente, está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la referencia correspondiente.

Lo anterior se entiende como complemento al Reglamento del Alumno de la Pontificia Universidad Católica de Chile (`http://admisionyregistros.uc.cl/alumnos/informacion-academica/reglamentos-estudiantiles`). Por ello, es posible pedir a la Universidad la aplicación de sanciones adicionales especificadas en dicho reglamento.