Código	Nombre
MDS7201	Proyecto de Ciencia de Datos

Nombre en inglés
Data Science Project

CT	Unidades Docentes	Horas de Cátedra	Horas Docencia Auxiliar	Horas de Trabajo Personal
6	10	3	0	7

Requisitos	Carácter del Curso
(CC5206/IN6531/MDS7102),(CC3201/MDS7103)	Obligatorio Magíster en Ciencia de Datos

Resultados de Aprendizaje

Este curso tiene como propósito que los estudiantes apliquen conceptos fundamentales de la Ciencia de Datos, así como obtener una visión general de lo que implica llevar a cabo un proyecto de datos. La finalidad es que los estudiantes resuelvan problemas mediante herramientas para el manejo, análisis y visualización de información. Por medio del desarrollo de un proyecto, el estudiante obtendrá conocimiento y dominio en el uso efectivo de herramientas de manejo y análisis de datos. Los proyectos serán problemas específicos definidos en diversos dominios de aplicación, donde los estudiantes deberán utilizar herramientas de análisis de datos y/o los lenguajes de programación que estimen pertinente (por ejemplo, Python, R, Octave, Weka, Knime y/u otros). Los problemas planteados estarán delimitados a la duración del curso en cuanto a su alcance y tamaño.

El proyecto seguirá la metodología estándar de un proyecto en ciencia de datos, compuesta por las siguientes etapas:

1. Comprensión y formulación del problema: se plantea el problema de ciencia de datos a abordar durante el proyecto.
2. Adquisición de datos: se identifican las fuentes de datos y se procede a extraer, limpiar y transformar dichas fuentes para su posterior análisis.

Al finalizar el curso, el estudiante:

3. Análisis y modelamiento de datos: se usan técnicas estadísticas, de minería de datos y de machine learning para extraer valor a partir de los datos con el fin de resolver el problema inicial.
4. Comunicación del resultado: se comunican los resultados usando técnicas de visualización de datos.
5. Despliegue: se pone en producción el modelo construido y validado.

En este curso el estudiante seguirá metodologías con estructura teórico-práctica, “hands-on”, donde se presenten cápsulas de contenido al inicio de la clase y luego se lleve a la práctica con ayuda del equipo docente. En este entorno, los estudiantes podrán ejercitar activamente, resolver dudas y fortalecer los conocimientos adquiridos en las cápsulas teóricas.

Es importante señalar que, en muchos casos, los modelos construidos con las técnicas vistas en este curso y a lo largo de todo el programa, se usan para la toma automática de decisiones.

Existe evidencia que estas decisiones pueden perjudicar a ciertos grupos demográficos cuando los datos poseen sesgos estereotipados (ej: las personas de cierta raza o género son más aptas para cierto tipo de trabajo). Esto se vuelve aún más complejo cuando los modelos carecen de interpretabilidad y transparencia. Durante el curso se dará un fuerte énfasis a estudiar este fenómeno y asegurar equidad y transparencia en los modelos diseñados en el proyecto.

Del mismo modo, el curso también pondrá especial cuidado en la privacidad de los datos, pues estos pueden contener información sensible (como orientación sexual o preferencia política) y consecuentemente deben ser utilizados con los fines específicos para los cuales fueron suministrados y no para propósitos arbitrarios

Los resultados de aprendizaje que se espera que el estudiante desarrolle son:

- Maneja bases de datos a través de los lenguajes de programación Python y R, considerando la revisión y limpieza de los datos, logrando adquirir conocimiento práctico en base al desarrollo de un proyecto de Ciencia de Datos.
- Analiza bases de datos para extraer valor a partir de estas, considerando actividades como minar datos, inferencia de datos y reconocimiento de patrones, en base al desarrollo de un proyecto de Ciencia de Datos.
- Trabaja en equipo en forma estratégica, colaborativa, responsable y organizada, considerando la autogestión y la autoevaluación de su desempeño, a fin de facilitar la innovación y la mejora en el desarrollo del proyecto de Ciencia de Datos.
- Comunica en forma oral y escrita los resultados del análisis de datos en forma eficiente y efectiva, a sus pares, académicos y otros profesionales.
- Cumple obligaciones y acuerdos, respetando los compromisos adquiridos, reflexionando sobre sus acciones y asumiendo las consecuencias.

Metodología Docente	Evaluación General
Este curso tiene una connotación práctica y de trabajo colaborativo. Se logra el aprendizaje mediante el desarrollo de un proyecto aplicado que involucra el uso intensivo de herramientas para el procesamiento, análisis y visualización de datos. Las cátedras serán enfocadas en la presentación de la problemática y la introducción de las herramientas de programación y técnicas de Ciencia de Datos, que podrán ser utilizadas por los equipos de trabajo durante el desarrollo de sus proyectos. Cada equipo de trabajo de 3 o 4 personas resolverá un problema de diferentes	El curso se evalúa a partir de hitos de avance del proyecto que miden el cumplimiendo de los objetivos propuestos al comienzo del semestre para los equipos de trabajo. Cada hito de avance puede corresponder a una presentación oral o a un informe escrito. Al finalizar el semestre, cada equipo deberá realizar una presentación oral de sus resultados y entregar un informe escrito con los resultados obtenidos. El cálculo de esas notas se efectúa de la siguiente forma: NH = Promedio de los hitos de avance. NP = Nota de presentación oral final.

disciplinas propuesto por el equipo docente. Cada problemática a resolver se presenta a los(as) estudiantes al inicio del curso, definiendo claramente el problema, los objetivos mínimos del proyecto y los datos pertinentes para la resolución del mismo. Cada equipo de trabajo contará con el apoyo de un integrante del equipo docente a lo largo del proyecto, con el fin de guiar el trabajo a realizar, apoyando en el uso de herramientas y la literatura. Además, periódicamente el equipo se reunirá con el/la investigador(a) que propone el problema para mostrar los avances del proyecto, y con ello facilitar la recepción de retroalimentación experta. Al finalizar el semestre, cada equipo de trabajo debe realizar una exposición oral final presentando el proceso a lo largo del semestre, las soluciones propuestas, limitaciones y líneas de investigación a seguir y resultados obtenidos en el proyecto. Finalmente, incluyendo los comentarios y correcciones resultantes de la presentación oral, el equipo de trabajo debe presentar el informe final escrito que resumen el trabajo realizado y resultados del proyecto.	NI = Nota de informe escrito final. Nota Final = 0,5 * NH + 0,25 * NP + 0,25 * NI La condición para aprobar el curso es: NH >= 4.0, NP >= 4.0, NI >= 4.0

Unidades Temáticas

Número	Nombre de la Unidad	Duración en Semanas
1	**Definición del problema de ciencia de datos**	3

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
1. Formulación del proyecto 2. Definición de los aspectos y desafíos técnicos del proyecto 3. Introducción a los desafíos de datos presentes en el proyecto 4. Presentación de la literatura básica de la problemática	El estudiante: 1. Identifica las etapas generales que deberá ejecutar en el desarrollo de sus proyectos de Ciencia de Datos, desde el planteamiento del problema hasta la evaluación de la solución. 2. Establece relaciones relevantes entre lo leído y otros conocimientos desde una perspectiva personal, académica y profesional.	[3] Cap 2, [5],[6]

5. Declaración de los objetivos y niveles de logro esperados al final del proyecto	3. Reconoce los aspectos específicos de la problemática abordada y los desafíos presentes desde el punto de vista de análisis, organización y procesamiento datos. 4. Reconoce los logros esperados y las dificultades que deberá abordar a lo largo del desarrollo del proyecto. 5. Intercambia con sus pares, profesores, otros profesionales y actores relevantes conocimientos en el ámbito de la Ciencia de Datos en la FCFM, ideas sobre diferentes desafíos profesionales en torno a los proyectos trabajados. 6. Cumple obligaciones y acuerdos, respetando los compromisos adquiridos en sus actividades académicas. 7. Participa en discusiones, respetando otros puntos de vista y entregando su visión sobre el tema. 8. Propone objetivos, desafiantes y claramente definidos, transmitiendo al equipo confianza y entusiasmo respecto de los logros alcanzados y los por lograr.	

Número	Nombre de la Unidad	Duración en Semanas
2	**Desarrollo de proyecto de ciencia de datos**	10

Contenidos	Resultados de Aprendizajes de la Unidad	Referencias a la Bibliografía
1. Presentación de metodologías de trabajo. a. Proceso de descubrimiento de conocimiento en bases de datos (KDD) b. Team Data Science Process	El estudiante: 1. Planifica su trabajo para cumplir con los hitos asignados en el proyecto. 2. Identifica la literatura y las herramientas necesarias para abordar los objetivos del proyecto. 3. Define una propuesta metodológica de trabajo en equipo, que implica todas las etapas de	[1-9]

2. Presentación de herramientas de organización de trabajo. a. R Notebooks b. Jupyter Notebooks c. Google Colab 3. Estudio de herramientas de exploración, manipulación y preprocesamiento de datos. a. Uso básico de R b. Librería tidyverse c. Reducción de dimensionalidad d. Clustering para exploración de datos e. Análisis de itemsets frecuentes y reglas de asociación 4. Visualización de datos a. Guías prácticas de uso de visualizaciones para generación de hipótesis b. Librería ggplot2 c. Herramientas de visualización de datos de alta dimensionalidad 5. Estudio de aprendizaje de máquinas a. Aprendizaje supervisado b. Aprendizaje no supervisado c. Librería scikit-learn 6. Ética en datos e inteligencia artificial.	desarrollo y presentación de resultados. 4. Comparte al equipo información, conocimientos y experiencias de forma clara y precisa para aportar al logro de los objetivos comunes. 5. Explica el problema abordado con un sentido colectivo estratégico y propósito del proyecto. 6. Analiza y justifica el uso de herramientas de análisis, procesamiento e inferencia considerando los objetivos declarados en el proyecto. 7. Elabora juicios acerca de la información contenida en los datos y de los supuestos que sustentan los métodos utilizados. 8. Propone un diseño experimental y define métricas concretas para medir el desempeño de los algoritmos utilizados. 9. Explora oportunidades e ideas novedosas que agreguen valor al problema abordado. 10. Selecciona las mejores herramientas y sus parámetros. 11. Visualiza los resultados y propone estrategias efectivas para presentarlos. 12. Discute y analiza los resultados obtenidos y los contrasta con los objetivos declarados. 13. Identifica y mitiga los posibles sesgos hacia algún grupo demográfico de los modelos desarrollados. 14. Reconoce cuando es necesario tener modelos transparentes y explicables y propone mecanismos para construirlos. 15. Evalúa en forma continua el cumplimiento de las metas y objetivos, en el contexto del trabajo en equipo, realizando ajustes oportunos en las actividades.	

a. Sesgo en machine learning. b. Métodos para reducir sesgo. c. Inteligencia artificial explicable. 7. Estudio de herramientas de estructura y procesamiento de datos especializadas 8. Privacidad: datos sensibles, derechos sobre datos y disponibilización. 9. Desarrollo del proyecto en base a hitos de avance 10. Presentación de la literatura básica de la problemática. 11. Declaración de los objetivos y niveles de logro esperados al final del proyecto.	16. Aprender a identificar datos personales, la propiedad y los derechos sobre los datos. 17. Entender las responsabilidades de las instituciones (o científicos de datos) sobre el manejo de los datos y cómo compartirlos con otros	

Número	Nombre de la Unidad	Duración en Semanas
3	**Presentación de solución**	2
Contenidos	**Resultados de Aprendizajes de la Unidad**	**Referencias a la Bibliografía**
1. Desarrollo de informes de resultados de un proyecto de Ciencia de Datos. 2. Presentación oral de un proyecto de Ciencia de Datos.	El estudiante: 1. Comunica de manera la solución propuesta para el problema dado. 2. Muestra el proceso completo dentro del cual se desarrolló el proyecto y las decisiones tomadas. 3. Cambia de registro de habla (formal, informal), académico, profesional, divulgativo) y los combina según las diferentes audiencias a las que se dirige.	[1-5]

Bibliografía General
[1] Grolemund, G; Wickham H. (2017). R for Data Science, Visualize, Model, Transform, Tidy and Import Data. O´Reilly. [En línea: https://r4ds.had.co.nz/index.html] [2] Kyrian Dale. (2016) “Data Visualization with Python and JavaScript: Scrape, Clean, Explore & Transform Your Data” [3] Shmueli, G., Patel, N. R., Bruce, P. C. (2010): Data Mining for Business Intelligence. 2nd ed., John Wiley and Sons, Hoboken, New Jersey [En línea: https://doc.lagout.org/] [4] Ian H. Witten and Eibe Frank. (2005) Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems) [5] O´Neil, C & Schutt, R. (2014) Doing Data Science: Straight Talk From de Frontline. Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. Third Edition. [En línea: https://share.nxtcloud.net/] [6] Baumer, Ben (2015). A Data Science Course for Undergraduates: Thinking With Data. Article in The American Statistician. [En línea: https://www.tandfonline.com/doi/full/10.1080/] [7] Leskovec J., Rajaraman A., and Ullman J. (2014) Mining of Massive Datasets, Second Edition. Cambridge University Press. [En línea: http://www.mmds.org/] [8] Field A., Miles J., and Field Z. (2012) Discovering Statistics Using R. SAGE Publishing. [9] James G., Witten D., Hastie T., and Tibshirani R. (2017) An Introduction to Statistical Learning with Applications in R. Springer. [10] Provost F., and Fawcett T. (2013) Data Science for Business. O’Reilly. [11] Paarsch H., Golyaev K. (2016) A Gentle Introduction to Effective Computing in Quantitative Research. The Mit Press. [12] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Limitations and Opportunities (2019). Versión online: https://fairmlbook. org/(19.05. 2019).

Vigencia desde:	2019
Elaborado por:	Martin Schaub, Jorge F. Silva, Richard Weber, Macarena Zapata, Felipe Bravo, Mauricio Quezada y Juan Manuel Barrios
Validado por:	Comité Académico Magíster en Ciencia de Datos