

PROGRAMA DE CURSO
PROYECTO DE CIENCIA DE DATOS

A. Antecedentes generales del curso:

Departamento	Ciencias de la Computación					
Nombre del curso	Proyecto de Ciencia de Datos					
Nombre del curso en inglés	Data Science Project					
Código	CC5214		Créditos	6		
Horas semanales	Docencia	3,0	Auxiliares	0	Trabajo personal	7,0
Carácter del curso	Obligatorio	0		Electivo de Especialización	X	
Requisitos	CC5206 Introducción a la Minería de Datos/ EL4106 Inteligencia Computacional/ IN6531 Introducción a la Minería de Datos/ MA5204 Aprendizaje de Máquinas/ AUTOR					

B. Propósito del curso:

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 lenguajes de programación que estime 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.

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 la cápsulas teóricas.

Las competencias específicas (CE) y genéricas (CG) a las que tributa el curso son:

CE3: Gestionar bases de datos utilizando modelos, lenguajes de consulta asociados, técnicas eficientes de acceso de datos y aplicación de políticas de seguridad, con la finalidad de obtener información relevante.

CE4: Extraer información relevante, utilizando el proceso de descubrimiento de conocimiento de datos.

CG1: Comunicación académica y profesional

Comunicar en español de forma estratégica, clara y eficaz, tanto en modalidad oral como escrita, puntos de vista, propuestas de proyectos y resultados de investigación fundamentados, en situaciones de comunicación compleja, en ambientes sociales, académicos y profesionales.

CG3: Compromiso ético

Actuar de manera responsable y honesta, dando cuenta en forma crítica de sus propias acciones y sus consecuencias, en el marco del respeto hacia la dignidad de las personas y el cuidado del medio social, cultural y natural.

CG4: Trabajo en equipo

Ejecutar con su equipo, de forma estratégica, diversas actividades formativas propuestas, considerando la autogestión de sí mismo y la relación con el otro, asumiendo diversos roles: de líder, colaborador u otros, según requerimientos y objetivos, sin discriminar por género u otra razón.

CG5: Innovación

Concebir ideas viables y novedosas que generen valor para resolver necesidades latentes, materializadas en productos, servicios o en mejoras a procesos dentro de un sistema u organización, considerando el contexto sociocultural y económico y los beneficios para el usuario.

C. Resultados de aprendizaje:

Competencias específicas	Resultados de aprendizaje
CE3	RA1: 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.
CE4	RA2: 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.
Competencias genéricas	Resultados de aprendizaje
CG4 - CG5	RA3: 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.
CG1	RA4: Comunica en forma oral y escrita los resultados del análisis de datos en forma eficiente y efectiva, a sus pares, académicos y otros profesionales.
CG3	RA5: Cumple obligaciones y acuerdos, respetando los compromisos adquiridos, reflexionando sobre sus acciones y asumiendo las consecuencias.

D. Unidades temáticas:

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
1	RA1- RA2-RA3-RA4-RA5	DEFINICIÓN DEL PROBLEMA DE CIENCIA DE DATOS	3 semanas
Contenidos		Indicador de logro	
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. 5. Declaración de los objetivos y niveles de logro esperados al final del proyecto.		El estudiante: 1. Identifica las etapas generales que deberán 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. Reconoce los aspectos específicos de la problemática abordada y los desafíos presentes del 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.	
Bibliografía de la unidad		[3] Cap 2, [5],[6]	

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
2	RA2-RA3-RA4-RA5	DESARROLLO DE PROYECTO DE CIENCIA DE DATOS	10 semanas
Contenidos		Indicador de logro	
1. Presentación de metodologías de trabajo. 1.1 Proceso de descubrimiento de conocimiento en bases de datos (KDD) 1.2 Team Data Science Process 2 Presentación de herramientas de organización de trabajo. 2.1 R Notebooks 2.2 Jupyter Notebooks 2.3 Google Colab 3 Estudio de herramientas de exploración, manipulación y preprocesamiento de datos. 3.1 Uso básico de R 3.2 Librería tidyverse 3.3 Reducción de dimensionalidad. 3.4 Clustering para exploración de datos. 3.5 Análisis de itemsets frecuentes y reglas de asociación 4. Visualización de datos 4.1 Guías prácticas de uso de visualizaciones para generación de hipótesis. 4.2 Librería ggplot2 4.3 Herramientas de visualización de datos de alta dimensionalidad. 5. Estudio de aprendizaje de máquinas. 5.1 Aprendizaje supervisado 5.2 Aprendizaje no supervisado 5.3 Librería scikit-learn		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 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 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.	

6. Estudio de herramientas de estructura y procesamiento de datos especializadas. 7. Desarrollo del proyecto en base a hitos de avance.	12. Discute y analiza los resultados obtenidos y los contrasta con los objetivos declarados. 13. Evalúa en forma continua el cumplimiento de las metas y objetivos, en el contexto del trabajo en equipo, realizando ajustes oportunos en las actividades.
Bibliografía de la unidad	[1-9]

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
3	RA3-RA4-RA5	PRESENTACIÓN DE SOLUCIÓN	2 semanas
Contenidos		Indicador de logro	
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 efectiva y eficiente 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 de forma estratégica según las diferentes audiencias a las que se dirige.	
Bibliografía de la unidad		[1-5]	

E. Estrategias de enseñanza:

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 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 grupo se reunirá con el/la investigador(a) que propone el problema para reportar 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.

F. Estrategias de evaluación:

El curso se evalúa a partir de hitos de avance del proyecto que miden el cumplimiento de los objetivos propuestos al comienzo del semestre para los grupos de trabajo. Cada hito de avance puede corresponder a una presentación oral o a un informe escrito. Al finalizar el semestre, cada equipo deberá entregar un informe escrito con los resultados obtenidos y realizar una presentación oral de sus resultados.

El cálculo de esas notas se efectúa de la siguiente forma:

- NH = Promedio de los hitos de avance.
- NP = Nota de presentación oral final.
- 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

G. Recursos bibliográficos:

- [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.

H. Datos generales sobre elaboración y vigencia del programa de curso:

Vigencia desde:	Primavera 2019
Elaborado por:	Martin Schaub, Jorge F. Silva, Richard Weber, Macarena Zapata, Felipe Bravo, Mauricio Quezada y Juan Manuel Barrios.
Validado por:	Jefe Docente - Departamento de Ciencias de la Computación
Revisado por:	Área de Gestión Curricular