PROGRAMA DE CURSO

Código	Nombre
IN7580	Introducción a la ciencia de datos
Nombre en Inglés	
Introduction of Data Science	

SCT	Horas de Cátedra	Horas Docencia Auxiliar	Horas de Trabajo Personal
6	3	1,5	1,5

Requisitos	Carácter del Curso
	Electivo TI

Competencias a las que tributa el curso

CE2	Identificar problemas y/ u oportunidades de negocios, considerando un análisis cualitativo y cuantitativo de la organización público o privada. Asimismo, diseñar y aplicar procesos de cambio e innovación al interior de la organización.
CE4:	Evaluar técnica y económicamente propuestas que generen soluciones de valor en una organización o empresa, reconociendo las ventajas competitivas del uso de la tecnología y alineándose a la lógica y necesidades del negocio.
CE5:	Elaborar, implementar y promover propuestas tecnológicas derivadas de una análisis continuo que conlleva el (re)diseño del negocio en donde se integran la gestión y las TIC's.
CG1:	Comunicar y argumentar en forma oral y escrita, propuestas y resultados de proyectos de negocio con TI, mediante el uso eficaz de técnicas de persuasión y de negociación, considerando los diferentes contextos y audiencias.
CG2:	Trabajar en equipos multidisciplinarios en diferentes situaciones, considerando el abordar, de forma crítica y autocrítica, las diferentes materias inherentes a su profesión, como también el asumir diferentes roles y tareas que implican colaboración y liderazgo.

Propósito del curso

El curso Introducción a la ciencia de datos tiene como propósito que el estudiante aplique herramientas y metodologías de *data science* a un problema o necesidad en la industria a fin de identificar y comprender un problema y sus variables, modelarlo y evaluar los resultados de la solución propuesta.

El estudiante identificará relaciones ocultas y patrones en los datos. Este curso presenta algunos temas avanzados de *data science* desde el punto de vista de sus aplicaciones más recientes, tales como la minería de textos y la minería de redes sociales, sus aplicaciones al diseño de sistemas de recomendación y a la minería de opiniones y algunas técnicas avanzadas no cubiertas tradicionalmente.

Resultados de Aprendizaje	**Competencia a la que tributa (CE–CG)**
RA1: Identifica, con su equipo, un problema de ciencia de los datos, aplicando metodologías de análisis (KDD o CRISP - DM), a fin de proponer una solución a una necesidad específica de una empresa o detectar patrones y relaciones ocultas en los datos.	CE2–CG2
RA2: Utiliza herramientas de análisis (Python, R o rapidminer) con sus respectivas metodologías, pre procesando datos y visualizando resultados, a fin de aplicarlas a problemas de data science.	CE5
RA3: Modela un problema de ciencia de los datos, considerando hipótesis de trabajo, variables, selección de atributos, a fin de proponer mejoras a las necesidades de una empresa, cuyos resultados explica de manera clara y coherente.	CE5–CG1
RA4: Determina la coherencia entre los resultados aportados por *data science* y la propuesta de solución a un problema de negocio, contrastando dichos resultados, a fin de evaluar y determinar puntos de mejora a la propuesta.	CE4

Metodología Docente	**Evaluación General**
La metodología se basa en: - Talleres Prácticos. - Laboratorios. - Proyecto Final.	La evaluación es de proceso y contempla: -Talleres Prácticos (60%). Los alumnos desarrollaran en clases tareas aplicando las técnicas entregadas en el curso. -Estudio Final (40%). Los estudiantes deberán escoger un problema y aplicar una o más modelos sobre datos propios para aportar una solución.

Unidades Temáticas

Número	RA al que tributa	Nombre de la Unidad	Duración en Semanas
1	RA1	Identificación de un problema de Ciencia de Datos	4

Contenidos	Referencias a la Bibliografía
1.1. Definiciones básicas (ciencia de datos y Data Scientists). 1.1.1. Evolución de la minería de datos. 1.1.2. Actualidad en Inteligencia de Negocios. 1.1.3. Ejemplos, contexto, aplicaciones. 1.1.4. Datos e información. 1.1.5. Metodologías KDD – CRISP DM. 1.1.6. Data Manipulation – Analytics – Visualization. 1.1.7. Aprendizaje supervisado/no supervisado. 1.2. Clasificadores y clustering. 1.2.1. Identificación de problemas. 1.2.2. Planteamiento de problema/solución. 1.2.3. Entendimiento del problema, datos y tipos de soluciones. 1.2.4. Recopilación/exploración de datos e información 1.3. Definición y selección de variables. 1.3.1. Identificación de problemas. 1.3.2. Planteamiento de problema/solución. 1.3.3. Comprensión del problema, datos y tipos de soluciones. 1.3.4. Recopilación/exploración de datos e información. 1.4. Definición y selección de variables.	(4) Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. "O'Reilly Media, Inc.". [1] I. Witten y E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 3d ed. Morgan Kaufmann, 2011.

Número	RA al que tributa	Nombre de la Unidad	Duración en Semanas
2	RA2	Pre-procesamiento de datos	2
Contenidos			**Referencias a la Bibliografía**
2.1. Trabajar con *missing values*. 2.2. Balanceo de clases. 2.3. Imputación de datos. 2.4. *Outliers*. 2.5. Transformación de datos. 2.6. Reducción de dimensionalidad.			(1) Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). Data Mining: Practical machine learning tools and techniques. 3era Edición. Morgan Kaufmann. (3) Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Número	RA al que tributa	Nombre de la Unidad	Duración en Semanas
3	RA2, RA3	Técnicas básicas de minería de datos	1
Contenidos			**Referencias a la Bibliografía**
3.1. Muestreo. 3.2. Training–validation–testing. 3.3. Sobreajuste. 3.4. Hold out y k-fold cross validation. 3.5. Medidas de evaluación de modelos (accuracy, precision, recall). 3.6. Regresión lineal y logística.			(1) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). An Introduction to Statistical Learning. Springer.

Número	RA al que tributa	Nombre de la Unidad	Duración en Semanas
4	RA2, RA3	Modelos de minería de datos	7
Contenidos			**Referencias a la Bibliografía**
4.1. Técnicas avanzadas de clasificación (SVM, NN, DT, RF, Naive-Bayes). 4.2. Minería de textos y de opiniones. 4.3. Clustering (Jerárquicos, K-Means, Fuzzy C-means, dbscan, optics). 4.4. Sistemas de recomendación. 4.5. Minería de redes sociales.			(1) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). An Introduction to Statistical Learning. Springer. (6) Murphy, K. P. (2015). Machine Learning: A Probabilistic Perspective. Machine Learning: A Probabilistic Perspective. The MIT Press.

Número	RA al que tributa	Nombre de la Unidad	Duración en Semanas
5	RA4	Interpretación de datos	1
Contenidos			**Referencias a la Bibliografía**
5.1. Visualización, data products y análisis de datos visuales. 5.2. Provenance, privacy, ethics and governance. 5.3. Tendencia de la minería de datos.			(1) Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). Data Mining: Practical machine learning tools and techniques. 3era Edición. Morgan Kaufmann.

Bibliografía General
Bibliografía obligatoria [1] I. Witten y E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 3d ed. Morgan Kaufmann, 2011. [2] James, Gareth, et al. An introduction to statistical learning. Vol. 6. New York: Springer, 2013. [3] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. [4] Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. "O'Reilly Media, Inc.". *Bibliografía complementaria*

[5] J. Hernández Orallo, M. J. Ramírez Quintana, C. Ferri, Introducción a la minería de datos, Pearson Prentice Hall, 2004
[6] Murphy, Kevin P. Machine Learning: a probabilistic perspective. MIT press, 2015.

.

Vigencia desde:	2017
Elaborado por:	Sebastián Ríos, Constanza Contreras, Jorge Retamales
Validado por:	Comité de docencia
Revisado por:	Andrea Matamoros (asesora curricular)

Anexo 1:

Reglas del curso
1. Se recibirán las tareas en U-cursos, donde se explique el trabajo realizado y sus conclusiones. Este resumen debe contener: Introducción, Problema, Técnica aplicada, Resultados y conclusiones. Los Anexos no se consideran parte del Resumen. 2. Toda copia literal del material del expositor, Internet o trabajos de otros alumnos, será calificada con la nota mínima 1.0. Si se detecta la copia a otro alumno, lo que incurre en una falta ética, el alumno corre el riesgo de reprobar automáticamente el curso.

Anexo 2: Propuesta de actividades

Semana	Actividad	Tema
1	Presentación general curso (Parte 0)	Explicación de objetivos y metodología del curso. Introducción general de los conceptos y los contenidos que se abordarán durante el semestre. Muestra de sus aplicaciones y utilidad en la Industria.
2	Parte 1	Conceptos claves sobre ciencia de los datos
3	Parte 2	Identificación y planteamiento de problemas
4	Parte 2	Definición de variables y selección de atributos
5	Parte 3	Missing values, outliers, balanceo de datos
6	Parte 3	Transformación de atributos, reducción de dimensionalidad
7	Parte 4	Técnicas básicas de minería de datos
8	Parte 5: Clasificación	Naive-Bayes, Decision Trees, Random Forest
9	Parte 5: Clasificación	Neural net, Support Vector Machine
10	Parte 5: Clustering	Jerarquico, K-Means, K-Mediods
V	Vacaciones mitad semestre	
11	Parte 5: Clustering	Fuzzy C-Means, dbscan, optics
12	Parte 5: Minería de Texto	Pre-procesamiento y clasificación
13	Parte 5: Minería de Texto	Semántica latente y análisis de sentimientos
14	Parte 5: Análisis de Redes Sociales	Grafos, aplicaciones, métricas de las redes sociales
15	Parte 5: Sistemas de recomendación	Tipos de recomendación
16	Parte 6: Otros Temas	Privacidad, visualización de datos